Open Research Online

The Open University's repository of research publications
and other research outputs

# Crystal structures of two nucleic acid-binding proteins

## Thesis

How to cite:

Törõ, Imre (2000). Crystal structures of two nucleic acid-binding proteins. PhD thesis The Open University.

For guidance on citations see FAQs.

© 2000 The Author

https://creativecommons.org/licenses/by-nc-nd/4.0/

Version: Version of Record

Link(s) to article on publisher's website:
http://dx.doi.org/doi:10.21954/ou.ro.0000e2e9

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data policy on reuse of materials please consult the policies page.

oro.open.ac.uk

UNRESTRICTED



# Crystal Structures of Two Nucleic Acid–binding Proteins

**Imre Törõ**

A thesis submitted in partial fulfilment of the requirements of the Open University for the degree of Doctor of Philosophy

**March 2000**

Sponsoring Establishment: The National Institute for Medical Research

Mill Hill, London

Collaborating Establishment: the European Molecular Biology Laboratory

Meyerhofstrasse 1, 69117 Heidelberg, Germany

AUTHOR NO: R7606594
DATE OF SUBMISSION: 29 MARCH 2000
DATE OF AWARD: 23 AUGUST 2000

# Abstract


## The Crystal Structure of S1 Nuclease from *Aspergillus oryzae*

S1 nuclease from *Aspergillus oryzae* is a glycoprotein of 32 kDa molecular weight. The protein has two enzymatic activities: it is an endo–exonuclease with high specificity for single stranded nucleic acids, and it has an additional 3'–nucleotidase activity. S1 nuclease is widely used in molecular biology as a single–strand specific nuclease due to its high stability and efficiency. It cleaves single–stranded regions of nucleic acids producing 5'–nucleotides without significant side–reactions. The crystal structure of S1 nuclease has been determined to 1.7 Å resolution by molecular replacement based on the known structure of P1 nuclease from *Penicillinum citrinum,* which has 49 % sequence identity compared to S1. The overall fold and the active site of S1 nuclease is basically identical to that of P1 nuclease, and also very similar to Phospholipase C from *Bacillus cereus* and alpha–toxin from *Clostridium perfringens.* The characteristic feature of this family of enzymes is a trinuclear zinc cluster in their active sites. A BLAST search in the sequence databases revealed several other protein sequences from bacteria, protozoa and plants possessing an approximately 30 % sequence identity compared to S1 nuclease, but showing an almost complete conservation of structurally and functionally important residues. Soaking and co–crystallisation experiments with substrate analogues have been carried out in order to obtain an enzyme–substrate complex. These efforts have not resulted in the structure determination of any complexes under crystallisation conditions: no binding of substrate has been observed. Nevertheless, an enzyme mechanism has been proposed based on structural data of S1 nuclease and nucleases with similar active sites.

## The Crystal Structure of an Sm–Related Protein from *Archaeoglobus fulgidus*

In eukaryotes Sm and Sm–like proteins are the core components of the small nuclear ribonucleoprotein particles (snRNPs), which are involved in a variety of functions

including rRNA processing, tRNA maturation and pre–mRNA processing. The Sm proteins are 70 to 120 amino acids long and share a common bi–partite signature sequence. The spliceosome, where the transesterification reaction of splicing occurs, is assembled by several snRNPs named after their constituting snRNA: U1, U2, U4, U5 and U6. An snRNA contains a short single stranded, uridine rich sequence motif, where the Sm proteins bind, but the three–dimensional arrangement of the Sm proteins and the mode of binding is unknown. In humans there are seven different canonical Sm proteins, which according to biochemical and electron microscopic studies seem to form a seven membered ring in vitro. Recently two crystal structures of human Sm protein dimers have been published.

Interestingly Sm–related protein sequences have been found in the available genomic database of various Archaebacteria based on sequence homology. In contrast with eukaryotes only one or two Sm–related protein sequences have been identified in one organism. Their function is currently unknown, since analogous pre–mRNA splicing does not occur in Archaebacteria. Two Sm–related proteins of *Archaeoglobus fulgidus* have been cloned and expressed as fusion proteins. One of them called AF–Sm2 has been crystallised utilising ammonium sulphate as precipitant and solved to 1.95 Å resolution by SIRAS using a single mercury derivative. AF–Sm2 crystallises in a hexagonal space group (P6) and contains one molecule per asymmetric unit. The 77 residue long protein has a very similar fold compared to the solved human Sm protein structures: a short N–terminal α–helix followed by a five stranded, strongly bent, U–shaped β–sheet resulting in a barrel–like overall fold. Six AF–Sm2 molecules form a ring in the crystal structure mediated by extensive hydrophobic and hydrogen–bonding interactions. Gel filtration experiments have indicated a pH dependence of oligomerisation in accordance with the crystallisation experiences. Currently the target of the Sm–related proteins of *Archaeoglobus fulgidus* and the stochiometry of oligomerisation *in vivo* is completely unknown.

# Acknowledgements

First of all I wish to thank Dietrich Suck, my supervisor at EMBL for his help, continued support and patience throughout my Ph.D. studies. I would like to thank past members of Dietrich's group, particularly Christophe Romier, who generously helped me at the very beginning of my studies and Hiang Teo Dreher, our former technician, for her valuable practical advice.

I am very grateful to Joachim Meyer and Claudine Mayer for their advice and stimulating discussions on theoretical aspects of my work.

The work presented in the second half of this thesis was a group effort with significant contributions of other members of our research group. The cloning of the AF–Sm2 gene was entirely done by Dr. Martin Dreher and the first expressions and purifications as well as the crystallisation of the protein was carried out by Hiang Teo Dreher. Without their fundamental contribution this thesis could not be submitted in the present form.

I am indebted to a great many people for their scientific and personal support at EMBL and at NIMR, UK over the last three an a half years.

A very special thank to my wife Réka for her personal support and encouragement during my Ph.D. studies.

# List of Contents

# Chapter 2: Structure determination of S1 nuclease from *Aspergillus oryzae*

## Chapter 3: Structure Analysis of S1 Nuclease

**Part B: The crystal structure of an Sm–related protein from *Archaeoglobus fulgidus***

## Chapter 6: Structure analysis of AF–Sm2 from *Archaeoglobus fulgidus*

# List of Tables

# List of Figures

# List of Abbreviations

| | |
|---|---|
| AMP | adenosine monophosphate |
| ATP | adenosine triphosphate |
| CCD | charge coupled device |
| CCP4 | Collaborative Computational Project Number 4 |
| CD | circular dichroism |
| DESY | Deutsches Electronen–Synchrotron |
| DNA | deoxyribonucleic acid |
| EDTA | ethylenediaminetetraacetic acid |
| EMBL | European Molecular Biology Laboratory |
| $F_c$ | calculated structure factor amplitude |
| $F_o$ | observed structure factor amplitude |
| HEPES | *N*–2–hydroxyethylpiperazine–*N'*–2–ethane sulfonic acid |
| HIV | human immunodeficiency virus |
| IPTG | isopropyl–β–*D*–thiogalactopyranoside |
| MAD | multiwavelength anomalous dispersion |
| MD | molecular dynamics |
| MIR | multiple isomorphous replacement |
| MLR | maximum likelihood refinement |
| MMA | methyl–mercury acetate |
| MR | molecular replacement |
| MS | mass spectometry |
| MW | molecular weight |
| NAG | *N*–acetyl–glucosamine |
| NCS | non–crystallographic symmetry |

| | |
|---|---|
| ORF | open reading frame |
| PAGE | polyacrylamide gel electrophoresis |
| PCR | polymerase chain reaction |
| PDB | protein data bank |
| PEG | polyethylene glycol |
| PLC | phospholipase C |
| Pu | purine base |
| r.m.s. | root mean square |
| RNA | ribonucleic acid |
| rpm | revolutions per minute |
| RRM | RNA recognition motif |
| $\sigma$ | standard deviation |
| SA | simulated annealing |
| SDS | sodium dodecylsulphate |
| SIRAS | single isomorphous replacement with anomalous scattering |
| $S_N1$ | first order nucleophile substitution |
| $S_N2$ | second order nucleophile substitution |
| snRNP | small nuclear ribonucleoprotein particle |
| Tris | tris–hydroxymethyl–amino methane |
| $V_M$ | volume to mass ratio |

# Part A: The crystal structure of S1 nuclease from *Aspergillus oryzae*

# Chapter 1

# Introduction

## 1.1 Introduction to nucleases

By definition nucleases are enzymes which cleave phosphodiester bonds producing either nicks in double stranded DNA, shorter nucleic acid segments or nucleotides. Nucleases can be classified in terms of their various biochemical or structural properties. The following categories are based on how nucleases act on various types of nucleic acids. They can act on double stranded substrates, like most of the restriction enzymes, while certain nucleases act only on single stranded nucleic acid showing no or only minimal affinity toward double stranded substrates. Single strand specific nucleases are actually members of the structure selective nucleases, which can recognise a certain structural feature or conformation of the nucleic acids, rather than their sequence (reviewed by Suck, 1998). Nucleases, which cleave a nucleotide from the ends of the nucleic acid substrate are called exonucleases, while nucleases capable of hydrolysing phosphodiester bonds within the sequence are termed endonucleases. Another way of classification can be based on the chemical nature of the sugar component in nucleic acids. Ribonucleases act preferentially on RNA, while deoxyribonucleases cleave mostly single or double stranded DNA, but there exist nucleases which do not discriminate between RNA and DNA. In cases where the final cleavage products are nucleotides one distinguishes between nucleases producing

3' or 5' nucleoside phosphates. An important group of nucleases are the site specific nucleases which require a short unique nucleotide sequence where they cleave. A good example is the restriction enzymes.

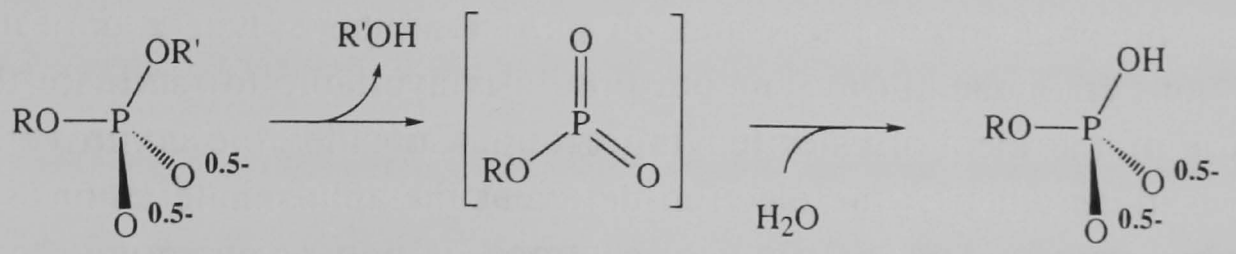
### 1.1.1 Phosphate ester hydrolysis

The common feature of nucleases is that they hydrolyse phosphodiester bonds. The phosphodiester bond, although it is *thermodynamically* unstable in an aqueous environment, is extremely stable even in 1 M NaOH due to the very high *kinetic* energy barrier to hydrolysis (Chin *et al.*, 1989). One reason for that is the negative charge of the phosphate group which causes electrostatic repulsion with the attacking hydroxide (Westheimer, 1987). The charge on the phosphate gives rise to $10^7$ times slower hydrolysis in the case of dimethyl phosphate compared to trimethyl phosphate (Guthrie, 1977). However, the speedup comparing enzyme catalysed hydrolysis by alkaline phosphatase to the spontaneous reaction, which actually hydrolyses phosphomonoester bonds, is approximately $10^{16}$ (Serpersu *et al.*, 1987). It is quite obvious that the active centre of the enzyme does more to accelerate the hydrolysis than simply neutralising the negative charge of the phosphate group.

Generally the enzyme catalysed phosphodiester bond cleavage proceeds with the cleavage of the P–O bond. Recently in a few cases another mechanism has been found which proceeds with the cleavage of the C–O bond. This ß–elimination reaction is exploited by repair enzymes acting on aldehydic abasic sites (Bailly & Verly, 1987; Mazumder *et al.*, 1990). Theoretically the hydrolysis of phosphate esters can proceed in two ways (Figure 1.1). One reaction mechanism is an $S_N1$ reaction: the prior dissociation of the leaving group forming a metaphosphate ester then addition of water. The other route is an $S_N2$ reaction: association with the attacking group forming a pentacoordinate

intermediate then the release of the leaving group. The current view nowadays based on theory and experimental results is that phosphodiester bond hydrolysis proceeds exclusively via the $S_N2$ mechanism with the exception of β–elimination mentioned above (Gerlt, 1992).

$S_N1(P)$




$S_N2(P)$

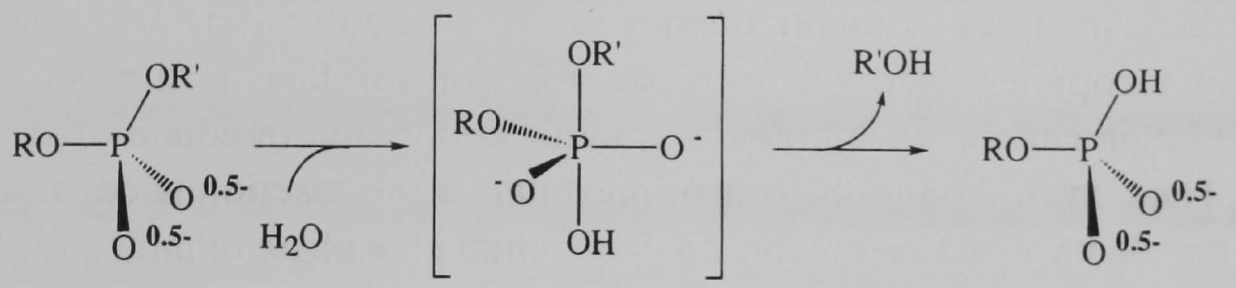



**Figure 1.1** The dissociative ($S_N1$) and associative ($S_N2$) mechanism of phosphodiester hydrolysis. For enzyme catalysed hydrolysis only the latter mechanism is relevant.

One consequence of the $S_N2$ mechanism is that it proceeds with the inversion of configuration. It means that the tetrahedrally coordinated reactant and the product are enantiomers if the ligands are all different but the attacking and leaving group are the same. However if the pentacoordinate intermediate has a long enough lifetime, in theory a reorganisation process, termed "pseudorotation" could occur which proceeds with the retention of configuration (Westheimer, 1968). Pseudorotation is a reorganisation of the pentacoordinate intermediate in a way that two equatorial ligands become axial and vice versa. While pseudorotation can occur in non–enzymatic displacement reactions of phosphate esters, up to now no enzyme catalysed phosphate ester hydrolysis proceeding with pseudorotation has been found.

## 1.1.2 The enzyme catalysed hydrolysis of phosphate esters

Biochemical experiments, especially those which focus on enzyme kinetics and reaction stereochemistry can provide a lot of useful information about the mechanism of nuclease action. However, the proper identification of the catalytic groups involved and the detailed reaction mechanism at the atomic level can be studied only if high resolution structural information is available. Fortunately, several nuclease structures have been solved at nearly atomic resolution, among them complexes with substrate, substrate analogue or inhibitor making it possible to draw some conclusions about their catalytic mechanism.

As was mentioned above the hydrolysis can proceed either with inversion of configuration on the phosphorous or with retention. The latter mechanism was found more infrequently compared to inversion of configuration. As the first step retention involves the displacement of the leaving group by a nucleophilic active site residue. In the second step the covalent intermediate is hydrolysed. A very good, well studied example, in which retention occurs, is alkaline phosphatase. As the first step of the hydrolysis, the residue S102 in the active centre forms a phosphate ester bond with the phosphate. This step proceeds with inversion of configuration. In the next step the phosphate is displaced by a zinc activated water, which step also involves inversion, therefore two inversions result in retention of configuration on the phosphorous (Figure 1.2; Holtz *et al.*, 1999). Another function of the active site residues may be to provide a general base which assists the attack of the nucleophile; a general acid which assists the dissociation of the leaving group and an electrophilic residue or residues which stabilise the pentacoordinate transition state by interacting with the negatively charged phosphate oxygen atoms (Mildvan, 1997). In addition to the active site residues metal ions can participate in the catalytic mechanism generally acting as Lewis acids. In the following paragraph selected examples with known

structure are discussed in relation to their catalytic mechanism.

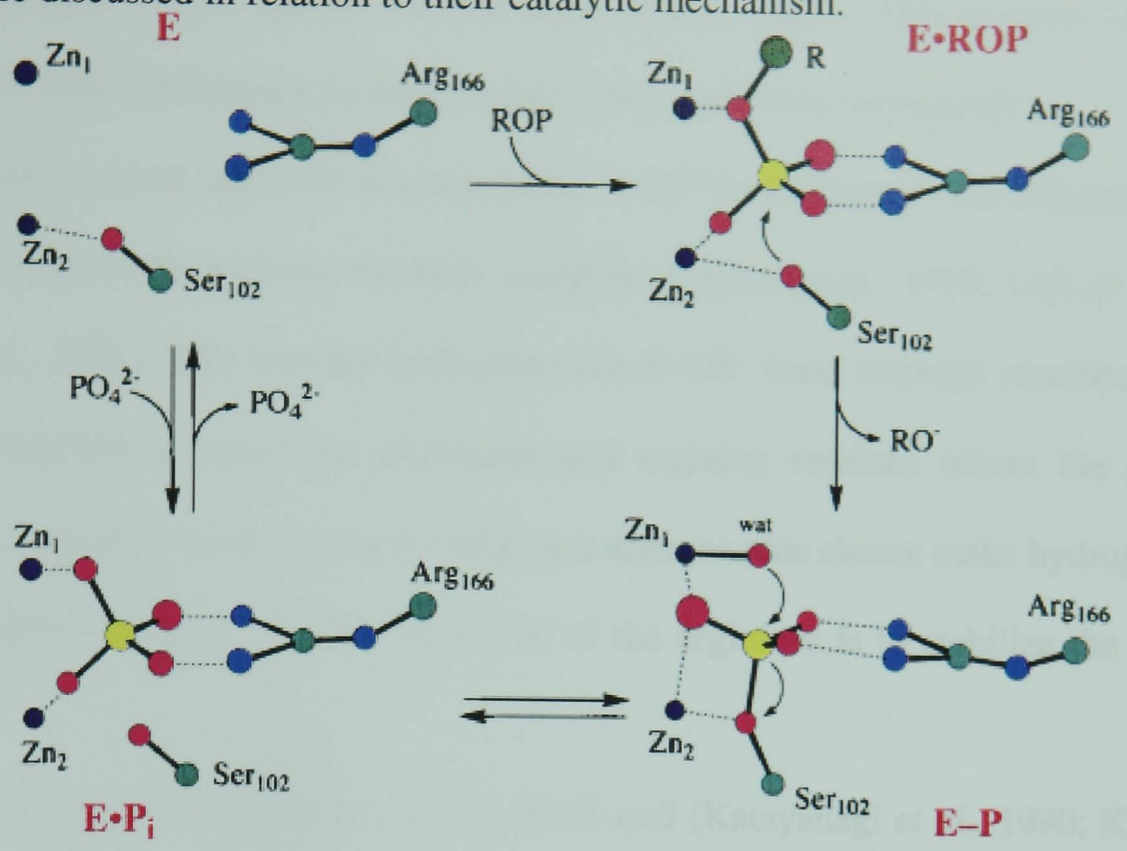



**Figure 1.2** The hydrolytic mechanism of alkaline phosphatase involves retension of configuration as a result of two times inversion ($S_N2$ reaction) (Holtz *et al.*, 1999).

### 1.1.3 The role of metal ions in the enzyme catalysed cleavage of phosphodiester bond.

Most of the nucleases depend on the presence of bivalent metal ions. They can participate in the catalytic mechanism acting as Lewis acids or simply stabilising the pentacoordinate intermediate by electrostatic interaction illustrated by the following examples.

In the Klenow fragment of *E. coli* DNA (Beese & Steitz, 1991; Brautigham & Steitz 1998) polymerase I, that possesses the 3'–5' exonuclease activity, magnesium and zinc play a multiple role in catalysis. One of the two closely spaced metal ions (Zn **A**) acts as a Lewis–acid for the attacking hydroxide and orientates a glutamic acid residue, which in turn acts as a base for the same hydroxide group. The other ion (Mg **B**) stabilises the leaving group and also stabilises the transition state directly by binding to one of the

phosphate oxygen atoms together with the other bivalent cation. This reaction mechanism, also found in other hydrolases, is the so called "two–metal ion mechanism".

Staphylococcal nuclease accommodates a $Ca^{2+}$ ion bound to the phosphate in the crystal structure of the enzyme inhibitor complex (Cotton *et al.*, 1979; Loll *et al.*, 1989; Hynes *et al.*, 1991). The enzyme hydrolyses the P–O5' bond through concerted general acid–base mechanism involving glutamate and arginine residues where the glutamate activates an attacking water molecule while two arginine side chains make hydrogen bonds to the phosphate oxygen. An additional role of the arginines is to stabilize the transition state.

The structures of ribonuclease H from *E. coli* (Katayanagi *et al.*, 1990; Katayanagi *et al.*, 1992) and the ribonuclease H domain of HIV–1 reverse transcriptase (Davies *et al.*, 1991) have been solved respectively to 1.48 and 2.4 Å resolution by X–ray crystallography. These ribonucleases degrade the RNA strand of a DNA–RNA hybrid in the presence of Mg by cleaving the P–O3' bond. Soaking the HIV ribonuclease H with $Mn^{2+}$ ions revealed two metal binding sites 4 Å apart coordinated by carboxylate groups (Davies *et al.*, 1991). These experimental results strongly suggest a catalytic mechanism very similar to the two metal ion mechanism described in the case of the Klenow fragment of DNA polymerase (Yang *et al.*, 1990), despite the fact that in *E. coli* ribonuclease H only one Mg binding site has been identified (Katayanagi e*t al.*, 1992).In fact, it has been shown for the Klenow fragment of *E. coli* DNA polymerase I that both metal ions are bound tightly only in the presence of substrate (Beese & Steitz, 1991).

Pancreatic bovine DNase I hydrolyses double stranded DNA by introducing nicks through cleavage of the P–O3' bond. It is neither sequence nor base specific, however the cleavage rate is strongly sequence dependent (Lomonosoff *et al.*, 1981; Drew & Travers 1984). The enzyme binds in the minor grove of dsDNA by widening and at the same time bending it towards the major groove (Suck *et al.*, 1988; Lahm and Suck, 1991;Lahm *et al.*, 1991). Site directed mutagenesis studies have shown that H134 and H252 are equally

important for catalysis (Doherty *et al.*, 1992, Worrall and Conolly, 1990). One of the DNaseI–dsDNA complexes solved to 2.3 Å resolution contains octamer dsDNA which is not cleaved by the enzyme (Weston *et al.*, 1992). The arrangement of the active site residues suggests a general acid–base mechanism. It has been proposed that H134 functions as a general acid, protonating the leaving O3', whereas H252 acts as a general base activating a water molecule by increasing its nucleophilicity. The metal ion, usually magnesium *in vivo*, is necessary to orientate the phosphate group and to stabilise the transition state.

S1 nuclease from *Aspergillus oryzae* (Ando, 1966), P1 nuclease from *Penicillinum citrinum* (Fujimoto *et al.*, 1974a) and *E. coli* endonuclease IV (Saporito & Cunningham, 1988) belong to the family of zinc dependent nucleases. Their structures have been solved (Volbeda *et al.*, 1991; Törõ & Suck, in preparation; Hosfield *et al.*, 1999) showing that all three nucleases possess a trinuclear zinc cluster in the active site, and probably share a common hydrolytic mechanism.

## 1.2 The role of zinc in the active site of zinc dependent hydrolases

There is a growing number of enzymes, including nucleases, which contain two or three catalytically important metal ions, very frequently zinc, in the active centre. The characteristic feature of this family of enzymes is a metal ion pair separated by about 3.5 Å, while sometimes a third metal ion is bound about 5 Å from the bimetal pair (reviewed by Wilcox, 1996; reviewed by Sträter *et al.*, 1996). To answer the question why zinc is so much favoured in comparison to other bi– or multivalent metal ions in protein structures, first one has to consider the electron configuration of the bivalent zinc ion (Berg and Shi, 1996). Since $Zn^{2+}$ has a completely filled *d*–shell it has no ligand field stabilization energy when coordinated by ligands in *any* geometry. For ions with partially filled *d* orbitals this

energy term can discriminate between various arrangements of the ligands. Another feature of the zinc ion is that according to hard–soft acid–base theory it is regarded as borderline acid. As a consequence, zinc can interact with a variety of ligands including sulphur from cysteine, nitrogen from histidine, lysine and the N–terminal amino group, with oxygen ligands of carboxylate groups and last but not least water. The electrochemical stability of $Zn^{2+}$ also makes it very suitable to play a structural or catalytic role in proteins. Under physiological conditions it is redox inactive: it can be neither reduced nor oxidized.

Vallee and Auld compared the X–ray structure of a dozen zinc dependent enzymes regarding the coordination of the catalytic zinc ions (Vallee & Auld, 1990a). Zinc forms complexes with nitrogen, oxygen and sulphur containing ligands having a binding frequency of His >> Glu > Asp = Cys. Water was found as a universal ligand and critical component of the catalytic site (Vallee & Auld, 1990a). Another interesting finding is that while the first two ligands are only 1–3 residues apart in the sequence, the third ligand is positioned considerably further, at least 19 residues from the second zinc coordinating residue. A similar study has shown a preference of ligands in the two cases when zinc plays a structural or catalytic role. Structurally important zinc is most frequently coordinated by four sulphur ligands while this is never the case for catalytic zinc (Vallee and Auld, 1990b). The same authors use the term "cocatalytic zinc binding site" in enzymes with two or three zinc atoms in close proximity to one another emphasizing their functional unity. A remarkable structural feature of these cocatalytic sites is the bridging carboxylate group of an aspartic or glutamic acid residue which bind two zinc ions (Vallee & Auld, 1993a; Vallee & Auld, 1993b).

A common feature in the mechanism of the bi– or trinuclear metallohydrolases, including zinc dependent nucleases, is the activation of a water molecule by a metal ion or ions acting as a Lewis acid. The water molecule, which should be considered rather as a hydroxide ion in such a case, is frequently forming a bridge between the two ions of the

above discussed "cocatalytic metal binding sites".

| *Enzyme* | *Abbreviation* | *Metal in the active site* | *Distance [Å]* | *Reference* |
|---|---|---|---|---|
| P1 nuclease | P1 | Zn1–Zn2–Zn3 | 3.2 (Zn1–Zn3) | Volbeda *et al.*, 1991 |
| S1 nuclease | S1 | Zn1–Zn2–Zn3 | 3.28 (Zn1–Zn3) | Törõ & Suck, in preparation |
| Endonuclease IV | – | Zn1–Zn2–Zn3 | 3.4 (Zn1–Zn2) | Hosfield *et al.*, 1999 |
| Phospholipase C | PLC | Zn1–Zn2–Zn3 | 3.3 (Zn1–Zn3) | Hough *et al.*, 1989 |
| Phosphotriesterase | PTE | Zn–Zn | 3.8 (Cd1–Cd2*) | Benning *et al.*, 1995 |
| Leucinaminopeptidase | LAP | Zn–Zn | 3.0 (Zn1–Zn2) | Sträter & Lipscomb, 1995 |
| Aminopeptidase from *A. proteolytica* | AAP | Zn Zn | 3.5 (Zn1–Zn2) | Chevrier *et al.*, 1994 |
| Bovine Calcineurin | PP–2B | Fe–Zn | 3.0 (Fe1–Zn2) | Griffith *et al.*, 1995 |
| Purple acid phosphatase | PAP | Fe–Zn | 3.1 (Fe1–Zn2) | Sträter *et al.*, 1995; Klabunde *et al.*, 1996 |
| Alkaline phosphatase | AP | Zn1–Zn2–Mg3 | 4.1 (Zn1–Zn2) | Kim and Wyckoff, 1991 |
| Ser/Thr phosphatase | PP–1 | Fe–Zn | 3.3 (Mn1–Mn2*) | Goldberg *et al.*, 1995 |
| DNA polymerase I (Klenow fragment) | Pol–I | Zn Mg | 3.9 (ZnA–MgB) | Beese & Steitz, 1991 |

* Other ion rather than the *in vivo* bound metal ion (Zn) was used for the structure determination

**Table 1.1** Zn–Zn (or Zn–other metal ion) distances in hydrolases with zinc ion(s) in the active site

The active site structure and catalytic mechanism of bi– or trinuclear metallohydrolases has been excellently reviewed by Wilcox (1996) and Sträter *et al.* (1996). Table 1.1 summarises the bi–or trinuclear zinc dependent hydrolases.

## 1.3 The single–strand specific nucleases S1 from *Aspergillus oryzae* and P1 from *Penicillinum citrinum*

So far, a number of single strand specific nucleases have been isolated from various sources, including fungi, bacteria, yeast, plants and animals. The best characterised single strand specific nucleases are: S1 nuclease from *Aspergillus oryzae*, Nuclease P1 from *Penicillinum citrinum*, mung bean nuclease I, *Neurospora crassa* nuclease, *Ustilago maydis* nuclease and BAL 31 nuclease. The first three nucleases are zinc dependent enzymes, *N. crassa* nuclease is an enzyme containing cobalt. Single strand specific nucleases are widely used in molecular biology to specifically cleave single stranded regions of double stranded nucleic acids leaving the double stranded regions mostly intact. This general feature of the enzymes can be exploited in a number of techniques as well as in the industrial preparation of mononucleotides from DNA and RNA. Due to their stability in a wide range of conditions and the absence of complicating side reactions P1 and S1 nuclease are the most frequently used single strand specific nucleases (Shishido & Ando, 1982; Fraser *et al.*, 1993).

The extracellulary secreted fungal proteins S1 nuclease from *Aspergillus oryzae* and P1 nuclease from *Penicillinum citrinum* have been isolated and characterised by Ando (1966) and Fujimoto (Fujimoto *et al.*, 1974a,b,c,d; Fujimoto *et al.*, 1975a,b). A nuclease in the *P. citrinum* cell culture was earlier reported to be useful for the preparation of 5'–mononucleotides from bulk RNA (Kuninaka, 1961).

The primary structure of S1 and P1 nucleases is known (Maekawa *et al.*, 1991;

Iwamatsu *et al.*, 1991). The comparison of the two sequences shows a sequence identity as high as 49% (Appendix A). The spread of identical residues is uniform all over the sequence suggesting also high structural homology. Besides their high sequence homology, their strikingly similar biochemical properties justifies discussing them together.

### 1.3.1 Biochemical properties of nuclease S1 and P1

The factors affecting the activity of S1 and P1 nucleases have been thoroughly studied and well described in the literature. They will be discussed below together with the substrate specificity of both nucleases.

S1 and P1 nucleases are heat stable enzymes. P1 nuclease is stable below 60 degrees whereas it shows highest activity at 70 degrees (Fujimoto *et al.*, 1974a,b). The heat stability of S1 nuclease has been utilised by one of the purification protocols heating to 75 degrees as a first step (Fujimoto *et al.*, 1974a; Vogt, 1973, Vogt, 1980). The stability at high temperature is believed to be the result of their high content of hydrophobic residues (over 50%) whereas the contribution of sugar to the stability at high temperature has been questioned by comparative analysis of the unmodified and glycosidase–treated enzyme (Shishido & Habuka, 1986).

The pH of highest enzymatic activity falls into the acidic range for both enzymes. It is pH 5.3 and 4.5 for P1 and S1 nucleases respectively. Elevating or decreasing the pH from the optimum results in a significant decrease of the enzyme activity (Fujimoto *et al.*, 1974b,c; Ando, 1966; Vogt, 1973). Ionic strength is another parameter which strongly affects the activity of S1 and P1 nucleases. For polyU and polyC as substrates P1 has a significantly decreased activity even at 200 mM NaCl concentration (Fujimoto *et al.*, 1974b). S1 nuclease also shows an activity optimum at 100 mM NaCl concentration,

having still 97% activity at 200 mM concentration, whereas 400 mM NaCl decreases its activity against ssDNA to 55% (Sutton, 1971).

Both nucleases are zinc dependent enzymes. Treatment with EDTA inactivates both enzymes in a stepwise manner which can be monitored by enzyme activity assays and CD–spectroscopy. Addition of EDTA to a concentration as high as 1 mM completely abolishes the activity of P1 that is reflected in a change of its CD spectra. The stepwise removal of zinc ions, eventually leading to unfolding of the enzymes, strongly suggests different roles of the individual zinc atoms bound in S1 and P1 nucleases. Addition of zinc partially restores activity and even 50 mM EDTA is insufficient to cause complete inactivation if added zinc is present. Three bound zinc atoms have been found in both nucleases as a result of atomic absorption spectroscopy and quantitative titration with EDTA monitoring conformational change by CD–spectroscopy. (Fujimoto *et al.*, 1974b; Fujimoto *et al.*, 1975a,b; Fujimoto *et al.*, 1980; Vogt, 1973; Shishido & Habuka, 1986).

The reported molecular weight of P1 and S1 nucleases are ~36 and 32 kDa respectively (Maekawa *et al.*, 1991; Vogt, 1973). However, there is a considerable difference between the molecular weight obtained by experimental methods (SDS–PAGE, MS, etc.) and the MW calculated on the basis of their sequences (29 kDa). The missing molecular mass can be attributed to glycosylated asparagine side chains. The sugar content has been analyzed for P1 nuclease by direct chemical methods. *D*–mannose, *D*–galactose and *N*–acetyl–glucosamine have been found in a ratio of 6:2:1. High binding affinity of P1 to concanavalin A Sepharose was also observed which indicates high mannose content (Fujimoto *et al.*, 1975a). Actually, one of the purification methods of S1 nuclease is based on its similar affinity to Con A Sepharose (Shishido & Habuka, 1986). The sequencing of S1 nuclease revealed the glycosylation sites N92 and N228 (Iwamatsu *et al.*, 1991).

## 1.3.2 Substrate specificity and inhibition of S1 and P1 nucleases

Both enzymes hydrolyse preferentially single–stranded DNA and RNA (Ando, 1966; Fujimoto *et al.*, 1974a). For S1 nuclease the cleavage rate for dsDNA has been found to be 75000 times lower in a comparative experiment (Wiegand *et al.*, 1975). The product of hydrolysis in both cases are 5'–mononucleotides (Ando, 1966; Fujimoto *et al.*, 1974a). P1 and S1 nucleases are exo–endonucleases: as the hydrolysis starts 5'–mononucleotides as well as shortened single–stranded fragments are detectable, which are finally hydrolysed to 5'–mononucleotides (Sutton, 1971). Besides the nuclease activity both nucleases possess an intrinsic 2'– and 3'–nucleotidase activity. The nucleotidase activity is lower in terms of reaction rates compared to the nuclease activity. Ribose–3'–phosphate and ribose–2'–phosphate are not cleaved. These findings strongly suggest that the minimal requirement of both nucleases is the presence of the base and a 3' (or 2') phosphate group in the substrate, however the type of the base slightly influences the cleavage rates as it was shown for 3'–mononucleotides and dinucleotides (Fujimoto *et al.*, 1974a,b,c,d; Oleson & Sasakuma, 1980; Oleson & Hoganson, 1981; Box *et al.*, 1993). There are quantitative differences between the specificity of the two nucleases. For P1 nuclease the relative rate of cleavage for different substrates is the following: RNA > ssDNA > 3'NMP > 3'dNMP > 2'NMP > dsDNA (Fujimoto *et al.* 1974c). In contrast, the best substrate of S1 nuclease is ssDNA, while RNA is cleaved with a two times lower rate. The nucleotidase activity of S1 is similar that of P1 nuclease (Oleson & Sasakuma, 1980; Oleson & Hoganson, 1981). The hydrolysis of the P–O3' bond proceeds with the inversion of configuration (Potter *et al.*, 1983a; Potter *et al.*, 1983b) which indicates that no covalent enzyme–substrate intermediate is involved in the hydrolytic mechanism.

As was found for other nucleases discussed in this chapter the products of hydrolysis may act as inhibitors. S1 nuclease is inhibited by various phosphate containing

compounds, like inorganic phosphate, pyrophosphate, 5' dAMP and 5'–dATP, the latter being the strongest inhibitor (Wiegand *et al.*, 1975; Oleson & Hoganson, 1981). Interestingly, while short oligonucleotides with 5'–phosphate (Fujimoto *et al.*, 1974d; Potter *et al.*, 1983b) have been found to be the best substrates for P1 nuclease, on the other hand the 5'–mononucleotides are the best inhibitors. Dinucleotides with 5' abasic nucleotide are not hydrolysed at all by P1 and S1 nucleases (Weinfield *et al.*, 1989). Dinucleotides with decreased aromaticity of the base of the 5'–nucleotide are, at best, weak substrates of P1 nuclease (Weinfeld *et al.*, 1993). These findings clearly indicate that the base of the 5'–nucleotide is crucial in the recognition of substrate, and additionally suggest the presence of an extended nucleotide binding site in the 5' direction from the catalytic site.

### 1.3.3 Action on double stranded nucleic acids

Single strand specificity does not mean that S1 and P1 nucleases do not cleave double stranded DNA at all. P1 or S1 nuclease introduce only a few nicks to dsDNA of phage ΦX174 (Godson, 1973) when compared to the contemporal total cleavage of single stranded DNA. Based on similar studies it has been proposed that zinc dependent single–strand specific nucleases cleave dsDNA at regions where single strands can form locally due to local melting or partial denaturation (St. John *et al.*, 1974; Wiegand *et al.*, 1975). Pulleyblank e*t al.* have proposed that the selectivity of zinc dependent single–strand specific nucleases is rather a consequence of the ability of these enzymes to recognise discrete conformations of the phosphodiester bonds that are simply rare in double stranded DNA, whereas more abundant in double stranded nucleic acid with non–A/non–B/non–Z conformations. Single stranded nucleic acids due to their higher intrinsic flexibility can adopt phosphodiester conformations recognised by these enzymes and therefore they are

cleaved more efficiently (Pulleyblank *et al.*, 1988). A good example is the GC/AT repeat which is a homopurine–homopyrimidin repeat with non–B/non–Z conformation and it is efficiently cleaved by S1 nuclease (Evans & Efstratiadis, 1986). S1 or P1 hypersensitivity can be caused also by chemical modifications like the loss of the base. dsDNA containing abasic sites on the opposite strand 1–3 bp apart is cleaved by S1 nuclease, while dsDNA with a discrete abasic site or two opposite abasic sites further apart than 3 bp is not sensitive to S1 nuclease, suggesting that the distortion in local conformation caused in the latter cases is not enough for the recognition by S1 nuclease. Single mismatches are also not sufficient for recognition and cleavage by S1 nuclease (Silber & Loeb, 1981). Hairpins also represent a "non–regular" conformation of dsDNA, and are more sensitive to single strand specific nucleases. P1 nuclease was reported to readily open hairpins leaving overhanging ends (Kabotyanski *et al.*, 1995).

### 1.3.4 The structures of P1 nuclease and its complexes with substrate analogues

The structure of P1 nuclease and its complexes with nuclease resistant oligonucleotide analogues have been solved (Lahm *et al.*, 1990; Volbeda e*t al.*, 1991, Romier *et al.*, 1998). The particular features of the structures are in good agreement with the results of biochemical experiments accumulated over two decades. The structure revealed three closely spaced zinc atoms bound in a cleft. Two of the zinc ions are only 3.2 Å apart, bridged by a water (or hydroxide ion) on the solvent side and by the carboxylate group of D120 from the protein's side. The third zinc ion is located about 5 Å away and has a different coordination sphere, suggesting, in accordance with the literature, a distinct catalytic function in the active site. Co–crystallisation and soaking of crystals with substrate analogues revealed a secondary nucleotide binding site with no clear functional

role. The crystal structures of P1 nuclease with substrate analogues either represent non–productive enzyme–substrate or enzyme–product complexes. A reaction mechanism has been proposed, which is in good accordance with the one proposed for the structurally very similar Phospholipase C from *Bacillus cereus* and endonuclease IV of *E. coli* (Hough *et al.*, 1989; Sundell *et al.*, 1994; Hosfield *et al.*, 1999). The latter enzyme is not homologous to P1 nuclease on the sequence level, however it has an active site with a strikingly similar trinuclear zinc cluster.

The structure of recombinant S1 nuclease has been recently solved, which is the subject of this part of the Ph.D. thesis and will be discussed in the following two chapters.

## 1.4 References

Ando, T. (1966). A nuclease specific for heat–denatured DNA isolated from a product of Aspergillus oryzae. *Biochim. Biophys. Acta* **114**, 158–168.

Bailly, V. & Verly, W. (1987). *Escherichia coli* endonuclease III is not an endonuclease but a β–elimination catalyst. *Biochem. J.* **242**, 565–572.

Beese, L.J. & Steitz, T.A. (1991). Structural basis for the 3'–5' exonuclease activity of *Escherichia coli* DNA polymerase I: A two metal ion mechanism. *EMBO J.* **10**, 25–33.

Benning, M.W., Kuo, J.M., Raushel, F.M. & Holden, H.M. (1995). Three–dimensional structure of the binuclear metal center of phosphotriesterase. *Biochemistry* **34**, 7973–7978.

Berg, J.M. & Shi, Y. (1996). The galvanization of biology: A growing appreciation for the roles of zinc. *Science* **271**, 1081–1085.

Box, H.C. et al. & Maccubin, A.E. (1993). The differential lysis of phosphoester bonds by nuclease P1. *Biochim. Biophys. Acta* **1161**, 291–294.

Brautigham, C.A. & Steitz, T.A. (1998). Structural Principles for the inhibition of the 3'–5' Exonuclease Activity of *Escherichia coli* DNA Polymerase I by Phosphorothioates. *J. Mol. Biol.* **277**, 363–377.

Chevrier, B., Schalk, C., D'Orchymont, H., Rondeau, J–M., Moras, D. & Tarnus, C. (1994). Crystal structure of Aeromonas proteolytica aminopeptidase: a prototypical member of the co–catalytic zinc enzyme family. *Structure*, **2**, 283–291.

Chin, J., Banaszczyk, F., Jubian, V. & Zou, X. (1989). Co(III) complex promoted hydrolysis of phosphate diesters: Comparison in reactivity of rigid cis–diaquotetraazacobalt(III) complexes. *J. Am. Chem. Soc.* **111**, 186–190.

Cotton, F.A., Hazen, E.E. & Legg, M.J. (1979). Staphylococcal nuclease: Proposed mechanism of action based on structure of enzyme–thymidine 3',5–biphosphate–calcium ion complex at 1.5 Å resolution. *Proc. Nat. Acad. Sci.* **76**, 2551–2555.

Davies, J.F., et al. & Matthews, D.A. (1991). Crystal Structure of the Ribonuclease H Domain of HIV–1 Reverse Transcriptase. *Science* **252**, 88–95.

Doherty, A.J., Worrall, A.F. & Connolly, B.A. (1992). Mutagenesis of the DNA binding residues in bovine pancreatic DNase I: an investigation into the mechanism of sequence discrimination by a sequence selective nuclease. *Nucleic Acids Res.* **19**, 6129–6132.

Drew, H.R. & Travers, A.A. (1984). DNA structural variations in the *E. coli* tyrT promoter. *Cell* **37**, 491–502

Evans, T. & Efstratiadis, A. (1986). Sequence–dependent S1 nuclease hypersensitivity of a heteronomous DNA duplex. *J. Biol. Chem.* **261**, 14771–14780.

Fraser, M.J. & Low, R.L. (1993). Fungal and mitochondrial nucleases. In *Nucleases*, 2nd edition, Linn, S.M., Lloyd, R.S. & Roberts, R.J., Eds., Cold Spring Harbor Laboratory Press, New York, pp. 171–207.

Fujimoto, M., Kuninaka, A. & Yoshino, H. (1974a). Purification of a nuclease from *Penicillinum citrinum*. *Agr. Biol. Chem.* **38**, 777–783.

Fujimoto, M., Kuninaka, A. & Yoshino, H. (1974b). Identity of phosphodiesterase and phosphomonoesterase activities with nuclease P1 (a nuclease from *Penicillinum citrinum*). *Agr. Biol. Chem.* **38**, 785–790.

Fujimoto, M., Kuninaka, A. & Yoshino, H. (1974c). Substrate specificity of nuclease P1. *Agr. Biol. Chem.* **38**, 1555–1561.

Fujimoto, M., Fujiyama, K., Kuninaka, A. & Yoshino, H. (1974d). Mode of action of nuclease P1 on nucleic acids and its specificity for synthetic phosphodiesters. *Agr. Biol. Chem.* **38**, 2141–2147.

Fujimoto, M., Kuninaka, A. & Yoshino, H. (1975a). Some physical and chemical

properties of nuclease P1. *Agr. Biol. Chem.* **39**, 1991–1997.

Fujimoto, M., Kuninaka, A. & Yoshino, H. (1975b). Secondary structure of nuclease P1. *Agr. Biol. Chem.* **39**, 2145–2148.

Gerlt, J.A. (1992). Phosphate ester hydrolysis. *Enzymes* **20**, 95–139.

Goldberg, J., Huang, H.B., Kwon, Y.G., Greengard, P., Nairn, A.C. & Kuriyan, J. (1995). *Nature* **376**, 745–753.

Godson, G.N. (1973). Action of the single–stranded DNA specific nuclease S1 on double–stranded DNA. *Biochim. Biophys. Acta* **308**, 59–67.

Griffith, J.P. *et al.* & Navia, M.A. (1995). X–ray structure of calcineurin inhibited by the immunophilin–immunosuppressant FKBP12–FK506 complex. *Cell* **82**, 507–522.

Guthrie, J.P. (1977). Hydration and dehydration of phosphoric acid derivetives: Free energies of formation of the pentacoordinate intermediates for phosphate ester hydrolysis and of monomeric metaphosphate. *J. Am. Chem. Soc.* **99**, 3991–4001.

Holtz, K.M., Stec, B. & Kantrowitz, E. (1999). A Model of the Transition State in the Alkaline Phosphatase Reaction. *J. Biol. Chem.* **274**, 8351–8354.

Hosfield, D.J., Guan, Y., Haas, B.J., Cunningham, R.P. & Tainer, J.A. (1999). Structure of the DNA repair enzyme endonuclease IV and its DNA complex: double–nucleotide flipping at abasic sites and three–metal–ion catalysis. *Cell* **98**, 397–408.

Hough, E. et al. & Derewenda, Z. (1989). High–resolution (1.5 Å) crystal structure of phospholipase C from *Bacillus cereus*. *Nature* **338**, 357–360.

Hynes, T.R. & Fox, R.O. (1991). The Crystal Structure of Staphylococcal Nuclease Refined at 1.7 Å Resolution. *Proteins* **10**, 92–105.

Iwamatsu, A., Aoyama, H., Dibó, G., Tsunasawa, S. & Sakiyama, F. (1991). Amino acid sequence of nuclease S1 from Aspergillus oryzae. *J. Biochem.* **110**, 151–158.

Kabotyanski, E.B., Zhu, C., Kallick, D.A. & Roth, D.B. (1995). Hairpin opening by single–strand–specific nucleases. *Nucleic Acids Res.* **23**, 3872–3881.

Katayanagi, M. et al. & Morikawa, K. (1990). Three–dimensional structure of ribonuclease H from *E. coli*. *Nature* **347**, 306–309.

Katayanagi, M. et al. & Morikawa, K. (1992). Structural Details of Ribonuclease H from *Escherichia coli* as Refined to an Atomic Resolution. *J. Mol. Biol.* **223**, 1029–1052.

Kim, E.E., Wyckoff, H.W. (1991). Reaction mechanism of alkaline phosphatase based on crystal structures. *J. Mol. Biol.* **218**, 449–464.

Klabunde, T., Sträter, N., Fröhlich, R., Witzel, H. & Krebs, B. (1996). Mechanism of Fe(III)–Zn(II) purple acid phosphatase based on crystal structure. *J. Mol. Biol.* **259**, 737–748.

Kuninaka, A., Kibi, M., Yoshino, H. & Sakaguchi, K. (1961). Studies on 5'–phosphodiesterase in microorganisms, Part II, Properties and application of Penicillinum citrinum 5'–phosphodiesterase. *Agr. Biol. Chem.* **25**, 693.

Lahm, A., Volbeda, A. & Suck, D. (1990). Crystallization and preliminary crystallographic analysis of P1 nuclease from *Penicillinum citrinum*. *J. Mol. Biol.* **215**, 207–210.

Lahm, A. & Suck, D. (1991). DNase I–induced DNA conformation: 2 Å structure of a DNase I–octamer complex. *J. Mol. Biol.* **222**, 645–667.

Lahm, A., Weston, S.A. & Suck, D. (1991). Structure of DNase I. *Nucleic Acids Molec. Biol.* **5**, 171–186.

Loll, P. & Lattman, E.E. (1989). The crystal structure of the ternary complex of staphylococcal nuclease, $Ca^{2+}$, and the inhibitor pdTp, refined at 1.65 Å. *Proteins Struct. Funct. Genet.* **5**, 183–201.

Lomonosoff, G.P., Butler P.J.G. & Klug, A. (1981). Sequence–dependent variation in the conformation of DNA. *J. Mol. Biol.* **149**, 745–760.

Maekawa, K., Tsunasawa, S., Dibó, G. & Sakiyama, F. (1991). Primary structure of nuclease P1 from Penicillinum citrinum. *Eur. J. Biochem.* **200**, 651–661.

Mazumder, A. *et al.* & Bolton, P.H. (1990). UV endonuclease V from bacteriophage T4 catalyses DNA strand cleavage at aldehydic abasic sites by a syn β–elimination

mechanism. *Biochemistry* **29**, 1119–1126.

Mildvan, A.S. (1997). Mechanism of signaling and related enzymes. *PROTEINS: Structure, Function and Genetics* **29**, 401–416.

Oleson, A.E. & Sasakuma, M. (1980). S1 nuclease of *Aspergillus oryzae:* A glycoprotein with an associated nucleotidase activity. *Arch. Biochem. Biophys.* **204**, 361–370.

Oleson, A.E. & Hoganson, E.D. (1981). S1 nuclease of *Aspergillus oryzae:* Characterization of the associated phosphomonoesterase activity. *Arch. Biochem. Biophys.* **211**, 478–481.

Potter, B.V.L., Connolly, B.A. & Eckstein, F. (1983a). Synthesis and configurational analysis of a dinucleoside phosphate isotopically chiral at phosphorus. Stereochemical course of Penicillinum citrinum nuclease P1 reaction. *Biochemistry* **22**, 1369–1377.

Potter, B.V.L., Romaniuk, P.J. & Eckstein, F. (1983b). Stereochemical course of DNA hydrolysis by nuclease S1. *J. Biol. Chem.* **258**, 1758–1760.

Pulleyblank, D.E., Glover, M., Farah, Ch. & Hanniford, D.B. (1988). In Wells. R.D. & Harvey, S.C. (eds), *Unusual DNA structures*. Springer, Heidelberg, pp. 23–44.

Rokugawa, K., Fujimoto, M., Kuninaka, A. & Yoshino, H. (1980). The role of zinc atoms in nuclease P1. *Agr. Biol. Chem.* **44**, 1987–1988.

Romier, C., Dominguez, R., Lahm, A., Dahl, O. & Suck, D. (1998). Recognition of single–stranded DNA by nuclease P1: High resolution crystal structures of complexes with substrate analogs. *PROTEINS:Structure, Function and Genetics* **32**, 414–424.

Saporito, S.M. & Cunningham, R.P. (1988). Nucleotide sequence of the nfo gene of *Escherichia coli* K–12. *J. Bacteriol.* **170**, 5141–5145.

Serpersu, E.H., Shortle, D. & Mildvan, A.S. (1987). Kinetic and magnetic resonance studies of active–site mutants of staphylococcal nuclease: Factors contrbuting to catalysis. *Biochemistry* **26**, 1289–1300.

Shishido, K. & Ando, T. (1982). Single–strand specific nucleases. In *Nucleases*, Linn, S.M. & Roberts, R.J., Eds., Cold Spring Harbor Laboratory Press, New York, pp. 155–185.

Shishido, K. & Habuka, N. (1986). Purification of S1 nuclease to homogeneity and its chemical, physical and catalytic properties. *Biochim Biophys. Acta* **884**, 215–218.

Silber, J.R. & Loeb, L.A. (1981). S1 nuclease does not cleave DNA at single–base mis–matches. *Biochim. Biophys. Acta* **656**, 256–264.

St. John, T, Johnson, J.D. & Bonner, J. (1974). Degradation of duplex DNA by S1 nuclease from Aspergillus. *Biochem. Biophys. Res. Commun.* **57**, 240–247.

Sträter, N., Klabunde, T., Tucker, P., Witzel, H. & Krebs, B. (1995). Crystal structure of a purple acid phosphatase containing a dinuclear Fe(III)–Zn(II) active site. *Science* **268**,

1489–1492.

Sträter, N. & Lipscomb, W.N. (1995). Transition state analogue L–leucinephosphonic acid bound to bovine lens leucine aminopeptidase: X–ray structure at 1.65 Å resolution in a new crystal form. *Biochemistry* **34**, 9200–9210.

Sträter, N., Lipscomb, W.N., Klabunde, T. & Krebs, B. (1996). Enzymatische Acyl– und Phosphoryltransferreaktionen unter Beteiligung von zwei Metallionen. *Angew. Chem.* **108**, 2158–2191.

Suck, D., Lahm, A. & Oefner, C. (1988). Structure refined to 2 Å of a nicked DNA octanucleotide complex with DNase I. *Nature* **332**, 465–468.

Suck, D. (1998). DNA Recognition by Structure–Selective Nucleases. *Biopolimers* **44**, 405–421.

Sundell, S., Handen, S. & Hough, E. (1994). A proposal for the catalytic mechanism in phospholipase C based on interaction energy and distance geometry calculations. *Protein Eng.* **7**, 571–577.

Sutton, W.D. (1971). A crude nuclease preparation suitable for use in DNA reassociation experiments. *Biochim. Biophys. Acta* **240**, 522–531.

Vallee, B.L. & Auld, D.S. (1990a). Active–site zinc ligands and activated $H_2O$ of zinc enzymes. *Proc. Natl. Acad. Sci. USA* **87**, 220–224.

Vallee, B.L. & Auld, D.S. (1990b). Zinc coordination, function, and structure of zinc enzymes and other proteins. (1990b). *Biochemistry* **29**, 5647–5659.,

Vallee, B.L. & Auld, D.S. (1993a). New perspective on zinc biochemistry: cocatalytic sites in multi–zinc enzymes. *Biochemistry* **32**, 6493–6500.

Vallee, B.L. & Auld, D.S. (1993b). Cocatalytic zinc motifs in enzyme catalysis. *Proc. Natl. Acad. Sci. USA* **90**, 2715–2718.

Vogt, V.M. (1973). Purification and further properties of single–strand–specific nuclease from *Aspergillus oryzae*. *Eur. J. Biochem.* **33**, 192–200.

Vogt, V.M. (1980). Purification and properties of S1 nuclease from *Aspergillus oryzae*. In *Methods. in Enzymology* (Grossman, L. & Moldave, K., Eds.), Vol **65**, pp. 248–255, Academic Press, New York.

Volbeda, A., Lahm, A., Sakiyama, F. & Suck, D. (1991). Crystal structure of Penicillinum citrinum P1 nuclease at 2.8 Å resolution. *EMBO J.* **10**, 1607–1618.

Weinfeld, M., Luzzi, M. & Paterson, M.C. (1989). Selective hydrolysis by exo– and endonucleases of phosphodiester bonds adjacent to an apurinic site. *Nucleic Acids Res.* **17**, 3735–3745.

Weinfeld, M., Soderlind, K.–J.M. & Buchko, G.W. (1993). Influence of nucleic acid base aromaticity on substrate reactivity with enzymes acting on single–stranded DNA. *Nucleic Acids Res.* **21**, 621–626.

Westheimer, F.H. (1987). Why nature chose phosphates. *Science* **235**, 1173–1178.

Westheimer, F.H. (1968). Pseudorotation in the hydrolysis of phosphate esters. *Accts. Chem. Res.* **1**, 70–78.

Weston, S.A., Lahm, A. & Suck, D. (1992). X–ray structure of the DNase I–d(GGTATACC)$_2$ complex at 2.3 Å resolution. *J. Mol. Biol.* **226**, 1237–1256.

Wiegand, R.C., Godson, G.N. & Radding, C.M. (1975). Specificity of the S1 nuclease from Aspergillus oryzae. *J. Biol. Chem.* **250**, 8848–8855.

Wilcox, D.E. (1996). Binuclear metallohydrolases. *Chem. Rev.*, **96**, 2435–2458.

Worrall, A.F. & Connolly, B.A. (1990). The chemical synthesis of a gene coding for bovine pancreatic DNase I and its cloning and expression in *Escherichia coli. J. Biol. Chem.* **265**, 21889–21895.

Yang, W., Hendrickson, W.A., Crouch, R.J. & Satow, Y. (1990). Structure of Ribonuclease H Phased at 2 Å Resolution by MAD Analysis of the Selenomethionyl protein. *Science* **249**, 1398–1405.

# Chapter 2

# Structure determination of S1 nuclease from *Aspergillus oryzae*

## 2.1 Introduction

This chapter describes the experimental and computational methods involved in the structure determination of S1 nuclease. The protein has been obtained from the laboratory of K. Kitamoto (University of Tokyo) in a crude lyophilised form. Conventional purification methods have resulted in satisfactory amounts of pure protein to carry out crystallisation experiments. Initial crystallisation conditions were quickly found utilising sparse matrix screens. The optimisation of initial conditions have given slowly growing, but well diffracting crystals. The phase problem was solved by molecular replacement using the already known structure of nuclease P1 from *Penicillinum citrinum*. The two proteins have 49% sequence identity suggesting a similar fold, and as a consequence one expects to get a clear molecular replacement solution. Actually, even one P1 molecule as a search model was sufficient to solve the structure of S1 nuclease by locating the positions of both molecules in the asymmetric unit. Molecular replacement (MR) was followed by density modification and a subsequent change of the P1 nuclease sequence to the correct sequence of S1 nuclease. Rigid body refinement, simulated annealing and group–based B–factor refinement were run as initial refinement steps using CNS. The refinement was completed using a combination of CNS and REFMAC, taking advantage of the strength of both programs. The structure determination process is outlined in Figure 2.1.

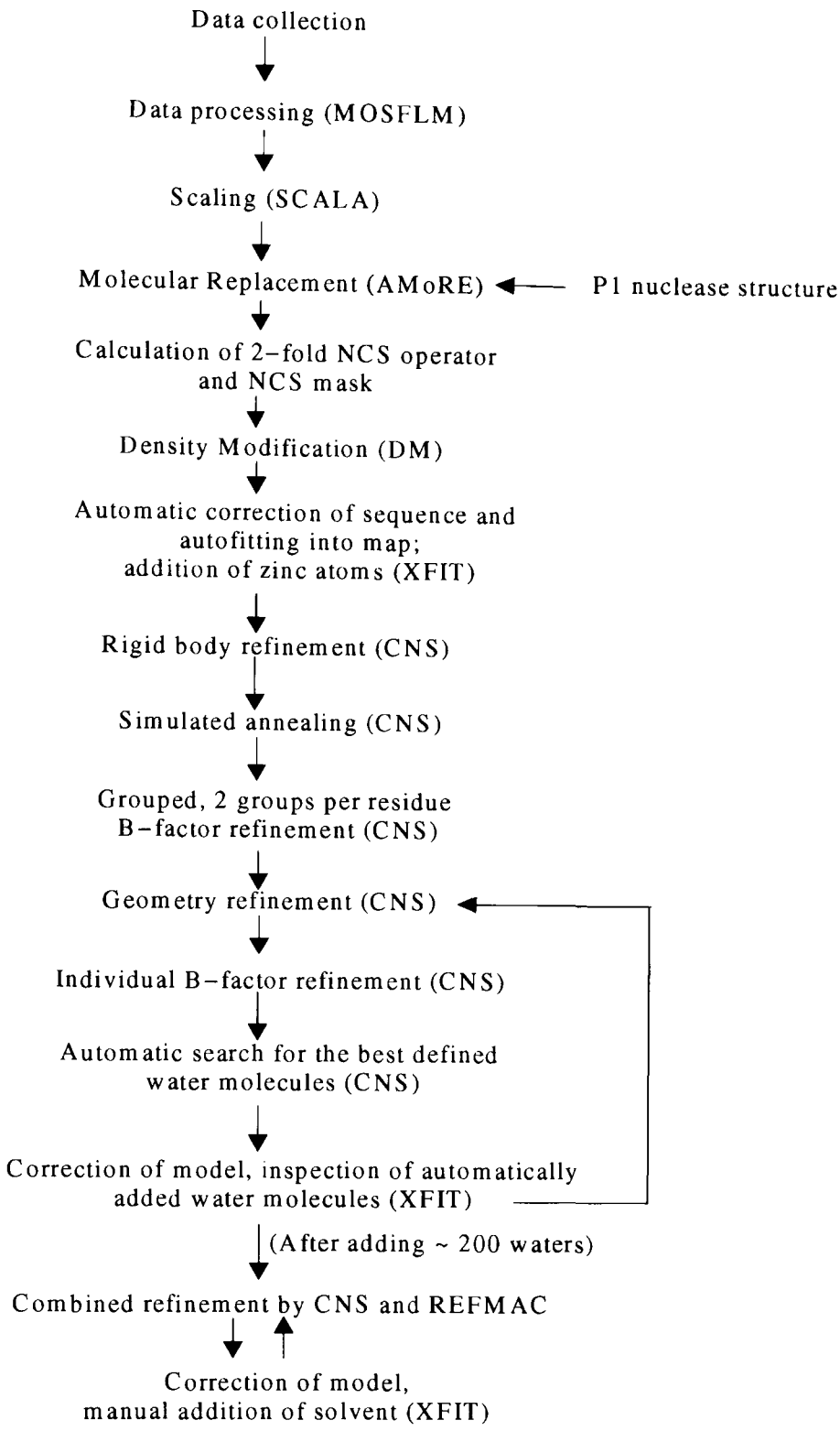



**Figure 2.1** Flow diagram of the steps involved in the structure determination of S1 nuclease. Most of the programs used here are part of the CCP4 Program Suite (CCP4, 1994), except MOSFLM (Leslie *et al.*, 1986), CNS (Brünger *et al.*, 1998) and XFIT (McRee, 1999).

## 2.2 Sample preparation and crystallisation

### 2.2.1 Expression and preliminary purification

Expression and preliminary purification was done in the laboratory of K. Kitamoto, University of Tokyo. The gene of S1 nuclease, *nucS* was cloned and placed under the control of the inducible strong promoter *glaA* on the expression plasmid. The recombinant protein was overexpressed in the native organism, *Aspergillus orizae* resulting in an approximately 1200–fold yield compared to the unmodified organism (Lee *et al.*, 1995). Since the overproduced protein is secreted to the extracellular space, the cells do not need to be lysed. The extract was concentrated using a 10 kDa ultrafiltration membrane. The concentrated sample was then treated with 80% saturated ammonium sulphate, and the precipitate was centrifuged. After centrifugation the precipitate was dissolved in S1 buffer (30 mM NaOAc, pH 4.6, 100 mM NaCl, 1 mM $Zn_2SO_4$) and dialysed against the same buffer. Following the ultracentrifugation of the dialysed protein solution, the sample was lyophilised and given to us for further purification and crystallisation.

### 2.2.2 Purification to homogeneity

400 mg of lyophilisate was dissolved in 20 ml of buffer A (30 mM NaOAc, pH 4.6, 1 mM $ZnCl_2$), and loaded on an equilibrated Q–Sepharose FastFlow column (Ø = 26mm, 40 cm long) from Pharmacia. The isoelectric point of the protein is around pH 4.3, thus at pH 4.6 it has a slightly negative net charge. The active group of the Q–Sepharose resin is a quaternary ammonium salt thus the negatively charged S1 molecules bind to the column at low salt levels. A linear salt gradient was applied to the column and the protein was eluted in ~300 mM NaCl. On the SDS–PA gel a tiny amount of impurity could be

detected, but its MW was much higher than the MW of S1 nuclease. The peak fractions were collected and concentrated on a 10 kDa ultrafiltration membrane, then loaded on a Superdex 75 HiLoad gel filtration column (Ø = 26mm). This final step resulted in a sufficiently homogeneous protein preparation (Figure 2.2). The peak fractions were dialysed against a storage buffer( 10 mM NaOAc, pH 4.6, 5 mM $ZnCl_2$ and 50 mM NaCl), then the pure protein was concentrated to a concentration of 20 mg/ml and 1.2 ml volume on an ultrafiltration membrane with 10 kDa cutoff. The concentration of the protein was measured on the basis of its absorption at 280 nm (Gill & von Hippel, 1989). The molar extinction coefficient of S1 nuclease at this wavelength is high due to its high content of aromatic amino acid side chains, thus low protein concentrations still give reliable absorbance values. The purified protein was analysed by mass–spectrometry. The spectrum (Figure 2.3) shows a series of peaks which are separated by molecular mass values typical of carbohydrate residues. Fortunately the different glycosylation states of the protein molecules did not prevent crystallisation.

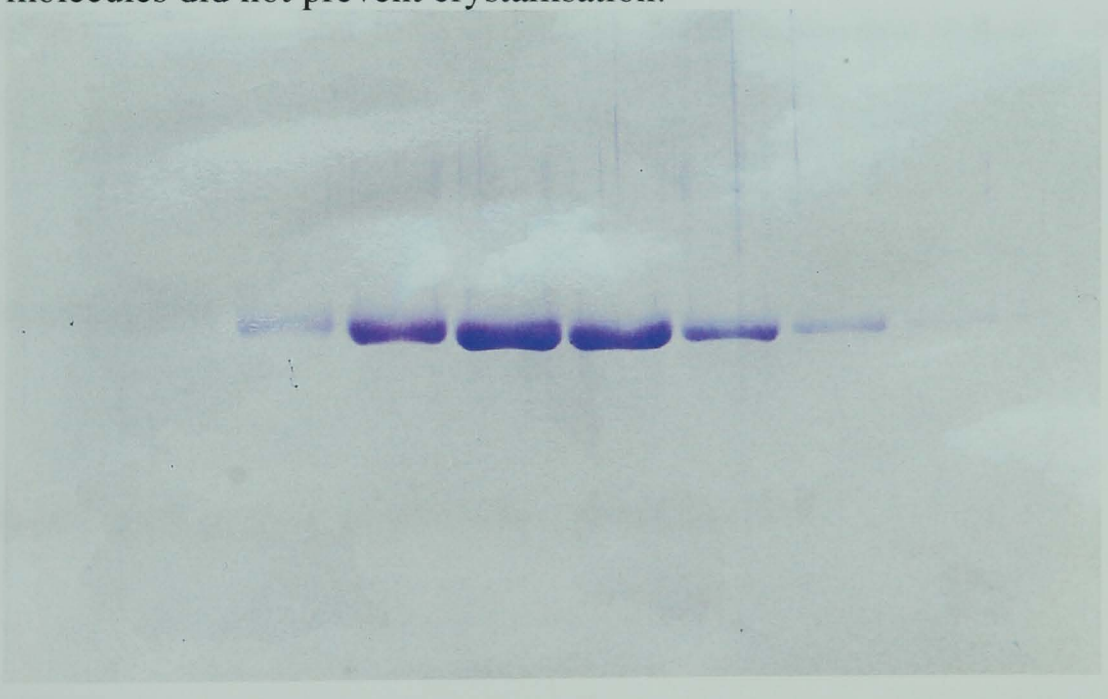


**Figure 2.2** SDS–PAGE of the S1 nuclease fractions after the final purification step (gel filtration). Each well of the gel was loaded with equal volume of eluate.

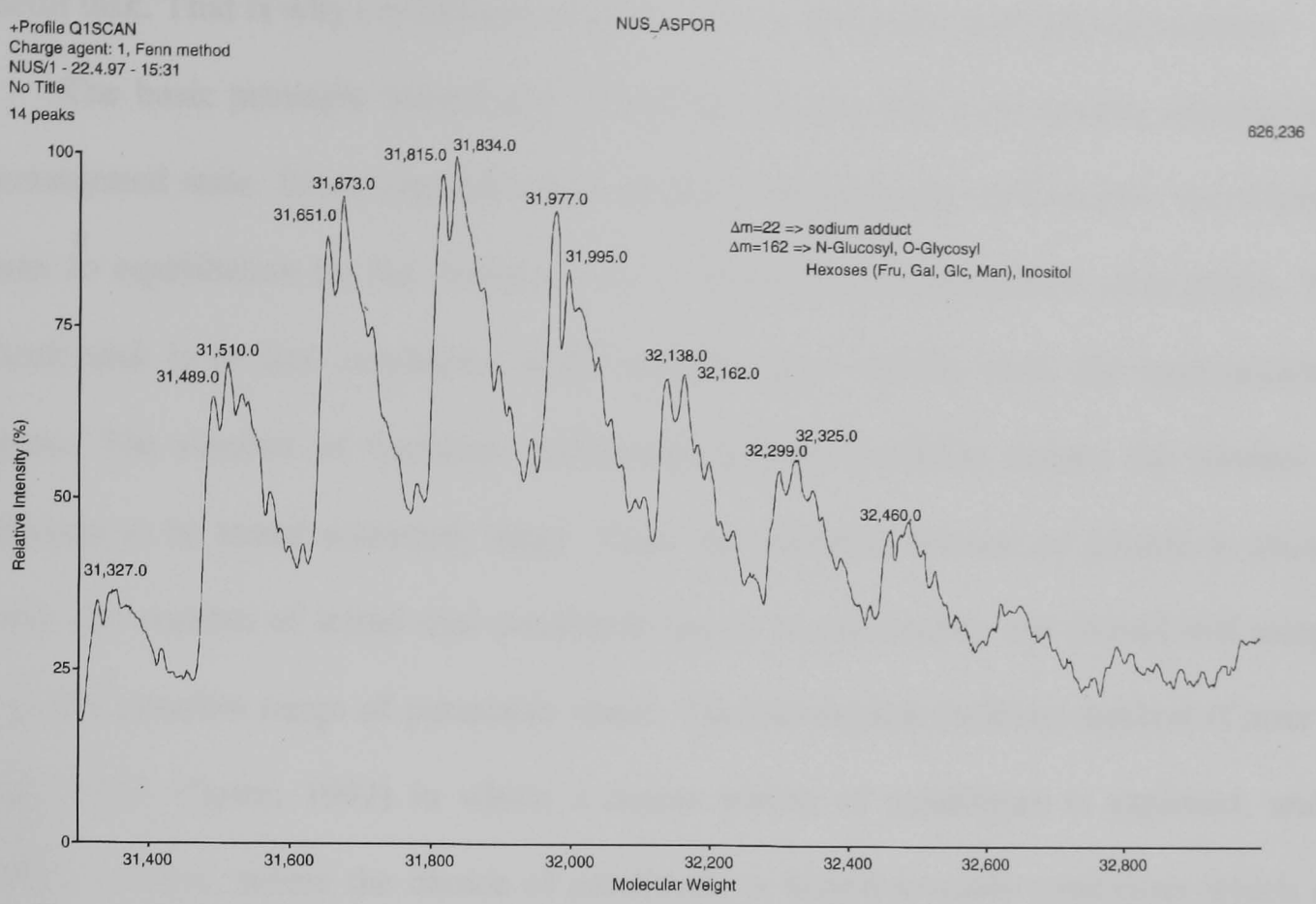



**Figure 2.3** Mass spectrum of purified S1 nuclease. The spectrum demonstrates the heterogeneity of the protein preparation due to different glycosylation level of the individual protein molecules. The mass difference between successive peaks corresponds approximately to the molecular weight of a hexose.

## 2.2.3 Crystallisation

### 2.2.3.1 Theory and praxis of protein crystallisation

Crystallisation is a crucial step in the process of structure determination by X–ray crystallography. It is also the least understood step despite the fact that the physical chemistry of crystal formation, nucleation, growth and cessation of growth have been extensively studied (McPherson, 1982; Fehér & Kam, 1985; Ducroix & Giegé, 1992). The large number of parameters which need to be explored in crystallisation experiments and the conformational flexibility of biomolecules, like proteins makes crystallisation a

difficult task. That is why crystallisation of proteins is still a trial and error procedure.

The basic principle behind any crystallisation is to bring the protein solution to a supersaturated state. Since supersaturation is thermodynamically unfavoured, the solution returns to equilibrium by the formation of a crystalline or amorphous solid phase. The difficult task is to find conditions where crystals grow slowly from the supersaturated solution. The number of variables influencing crystal formation makes the number of conditions to be tested extremely large. Since the available amount of protein is usually limited, the number of initial trial conditions has to be reasonable, but should still sample the widest possible range of parameter space. The incomplete factorial method (Carter & Carter, 1979; Carter, 1992) in which a coarse matrix of conditions is explored, and a modified version, where the choice of conditions is biased towards conditions which are already known to yield crystals (Jancarik & Kim, 1991), may give a good starting point for crystallisation. A careful optimisation of the parameters usually results in better quality crystals.

Supersaturation can be achieved by using various precipitants. A common method is to increase the effective concentration of the protein by adding salt or PEG (McPherson, 1985). A second method is to decrease the repulsive forces between the molecules by decreasing the ionic strength or by adding organic solvents which increase electrostatic interaction between the molecules (Blundell & Johnson, 1976). Two parameters from the large number of variables are usually the most influential in crystal growth: the effective pH and the temperature. A high level of purity of the protein sample is mandatory for successful crystallisation. The preparation has to be not only chemically pure, i.e. no contamination should be present, but the molecules have to be homogeneous at the molecular level. It means that the protonation state, disulphide bridges, and posttranslational modifications have to be the same in all the molecules.

In order to achieve supersaturation, several methods can be used. Popular methods are vapour diffusion, microbatch methods, dialysis and free interface diffusion (Ducroix &

Giegé, 1992). The most popular technique is vapour diffusion where a drop of the protein solution (usually 1 µl) is mixed with the same volume of well solution which has a significantly higher volume compared to the drop.

### 2.2.3.2 Crystallisation of S1 nuclease

The initial screens were set up using a protocol (Zeelen *et al.*, 1994) based on the incomplete factorial screen of Kim and Jancarik (Jancarik & Kim, 1991). The screening procedure was implemented as a pipetting robot program using 25 different solutions for the preparation of 48 well solutions. The robot consists of a standard Gilson autosampler and a motor–driven syringe. The control software (Oldfield *et al.*, 1991) of the robot makes it easily possible to modify the well solutions at will.

The screen quickly showed that the protein is prone to crystallise with PEG as precipitant and that the pH should be between 6.5 and 8.0. As a result of using 15–20 PEG2000 monomethylether as precipitant and a pH around 7.5 it was possible to obtain long clustered needles. Varying the average molecular weight of PEG and the pH, and trying several additives did not improve the quality of the crystals. Several crystallisation trials were set up utilising higher PEG concentrations (25%) and these resulted in nicely formed individual crystals. The interesting point in the crystallisation process was that the drops contained a voluminous amorphous precipitate after setting them up, while 1–2 months later the drops cleared up and a shower of nicely shaped crystals appeared (Figure 2.4). After optimisation the following conditions were found to be optimal for growth of this crystal form of S1 nuclease: 25% PEG2000 monomethylether, 100 mM HEPES, pH 8.0, 5 mM $ZnCl_2$. Despite the fact that many crystals grew in the drop it was possible to find crystals with a size of up to 0.25 mm as their longest dimension.

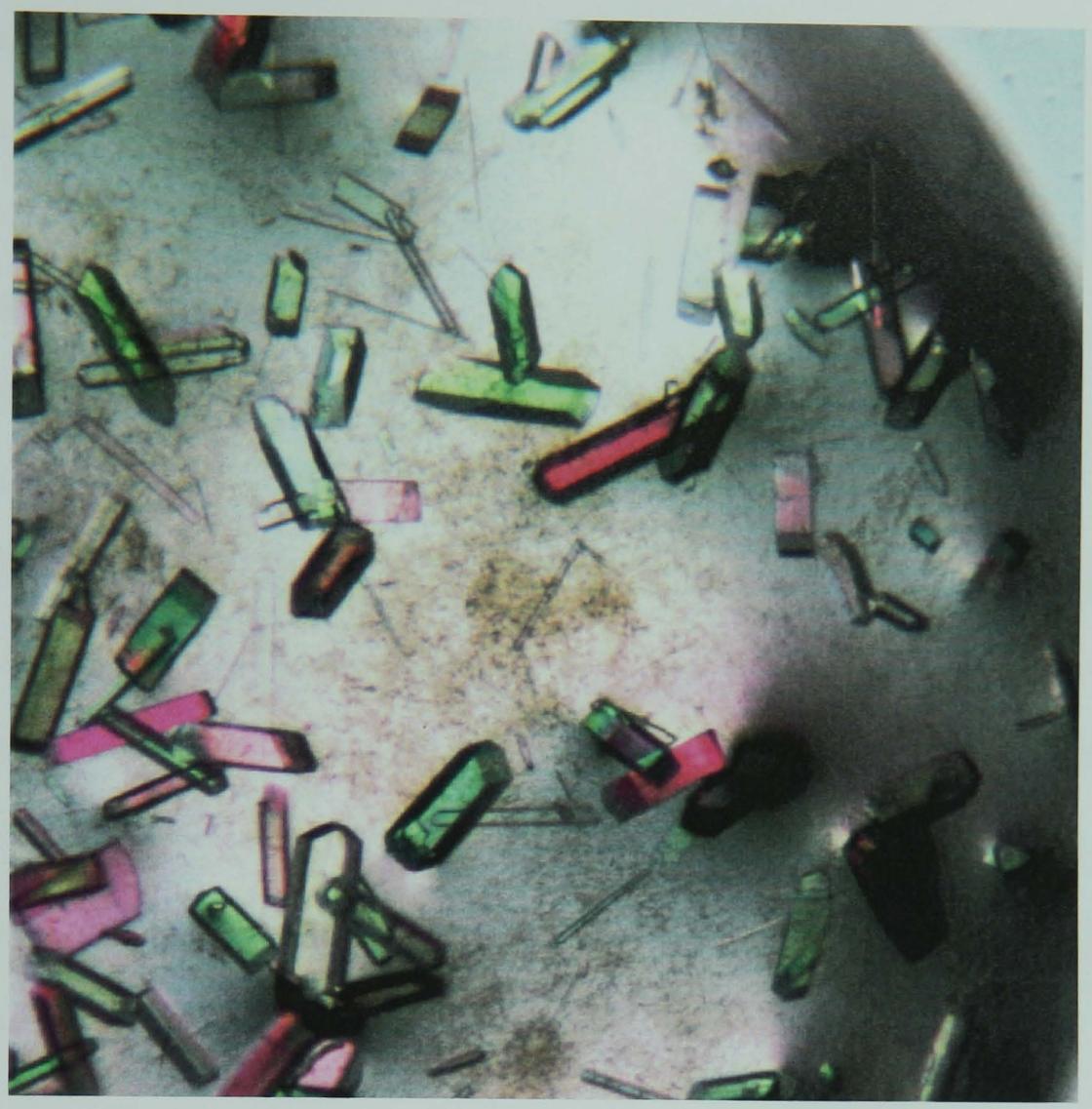


**Figure 2.4** S1 nuclease crystals growing from amorphous precipitate. In the drop shown here most of the precipitated protein is incorporated into crystals.

## 2.3 Data collection and processing

### 2.3.1 Introduction to diffraction theory

X–ray data collection from a protein crystal is the measurement of intensities diffracted by the crystal lattice. The detection methods can be either film, single photon counter, image plate, area detectors or charge coupled device camera (CCD). Nowadays, the most widely used detector types are the image plates and the CCD cameras (Gruner & Ealick, 1995) combined with the oscillation method that implies the rotation of the crystal around a single axis perpendicular to the beam (Arndt & Wonacott, 1977). After a certain

rotation range the detector device is read out and the diffraction pattern is digitally stored for analysis.

The scattering of X–rays on crystals is a result of the interaction between the X–rays as electromagnetic waves and the electrons of the molecules in the crystal. The waves scattered by the crystal are the vectorial sum of all waves each scattered by a single electron. The symmetry imposed by the packing of the individual molecules in the crystal lattice is reflected in the symmetry of the diffraction pattern which can be explained most simply by using the concept of a reciprocal lattice. The reciprocal lattice rotates exactly as the real lattice does. The observed reflection pattern can be easily constructed if we consider the condition of reflection, when a reciprocal lattice point is passing through a sphere with a radius of $1/\lambda$. This geometrical representation was proposed by Ewald, and is known as Ewald–construction (Figure 2.5).

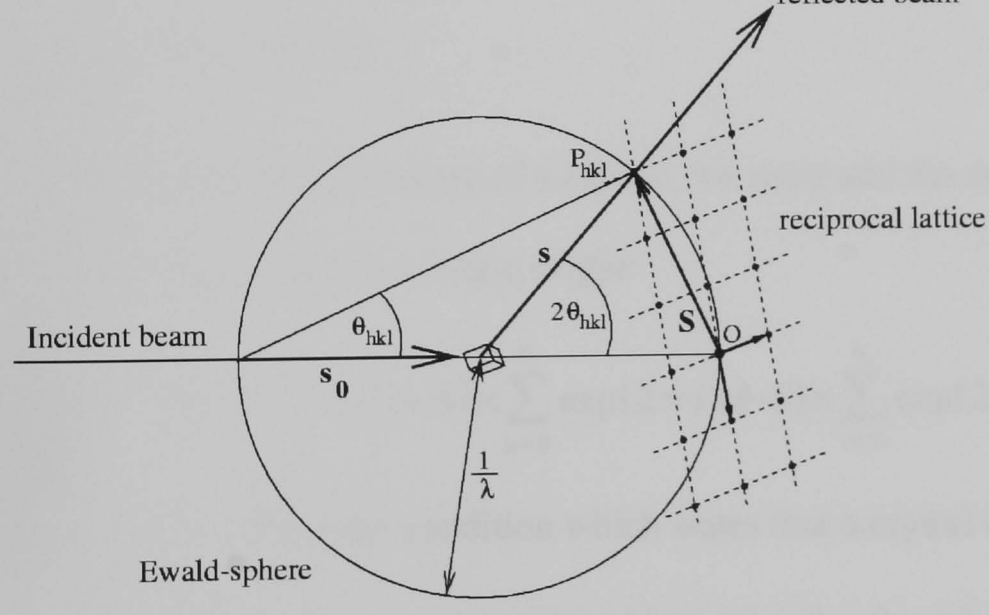



**Figure 2.5** The Ewald construction. *O* is the origin of the reciprocal lattice, $P_{hkl}$ is a reciprocal lattice point. The radius of the circle is $|s_0| = 1/\lambda$. Scattering occurs when the vector ***S*** has its endpoint $P_{hkl}$ on the sphere. The direction of scattering is ***s***.

While one can easily measure the intensity (*amplitude*) of the scattered beams by using a detection method mentioned above, it is impossible to measure the *phases* which have to

be obtained indirectly. Once the phases are available the electron density can be derived from the experimental intensities by the procedure of Fourier–transformation. Without proving them, some of the fundamental equations are presented below:

- The atomic scattering factor for a single atom is the function of the electron density:

$$f=\int_r \rho(\boldsymbol{r})\cdot\exp(2\pi i\boldsymbol{r}\cdot\boldsymbol{S})d\boldsymbol{r}$$ , where $\boldsymbol{S}$ is the vector $\boldsymbol{s}-\boldsymbol{s_0}$ (Figure 2.5) and $\boldsymbol{r}$ is the vector pointing to the scatterer (the electron) from the origin, which is the nucleus in this case. $\rho(\boldsymbol{r})$ is the electron density at the end point of $\boldsymbol{r}$. The electron cloud of atoms is assumed to be spherically symmetric, therefore $f$ is always *real*:

$$f=2\int_r \rho(\boldsymbol{r})\cdot\cos(2\pi\boldsymbol{r}\cdot\boldsymbol{S})d\boldsymbol{r}$$ .

- If we consider an entire unit cell then we have to sum the atomic scattering as vectors which gives us the so called structure factor:

$$\boldsymbol{F}(\boldsymbol{S})=\sum_{j=1}^{n} f_j\cdot\exp(2\pi i\boldsymbol{r}\cdot\boldsymbol{S})$$ .

- Moreover, if we calculate the scattering of a crystal, we must add the scattering of each individual unit cell with respect to a single origin:

$$\boldsymbol{K}(\boldsymbol{S})=\boldsymbol{F}(\boldsymbol{S})\times\sum_{t=0}^{n_1}\exp(2\pi i t\boldsymbol{a}\cdot\boldsymbol{S})\times\sum_{u=0}^{n_2}\exp(2\pi i u\boldsymbol{b}\cdot\boldsymbol{S})\times\sum_{v=0}^{n_3}\exp(2\pi i v\boldsymbol{c}\cdot\boldsymbol{S})$$ .

The above equation leads to the Laue condition which states that a crystal can only diffract if the numbers $\boldsymbol{a}\cdot\boldsymbol{S}$ , $\boldsymbol{b}\cdot\boldsymbol{S}$ , $\boldsymbol{c}\cdot\boldsymbol{S}$ are integer numbers ($h$ $k$ $l$), otherwise the sum of the vectors would be equal to zero. Here $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ are the translation vectors in the crystal. The origin of the unit cells constituting the crystal lattice are at $t\cdot\boldsymbol{a} + u\cdot\boldsymbol{b} + v\cdot\boldsymbol{c}$, where $t$, $u$ and $v$ are integer numbers.

- The structure factor can be expressed as a function of the electron density:

$$\boldsymbol{F}(\boldsymbol{S})=\int_{cell}\rho(\boldsymbol{r})\cdot\exp(2\pi\boldsymbol{r}\cdot\boldsymbol{S})d\nu$$ , where $d\nu=V_{cell}\cdot dxdydz$ ($x$, $y$, $z$ are now fractional coordinates).

- Since $\boldsymbol{r}\cdot\boldsymbol{S}=(\boldsymbol{a}\cdot x+\boldsymbol{b}\cdot y+\boldsymbol{c}\cdot z)\cdot\boldsymbol{S}=\boldsymbol{a}\cdot\boldsymbol{S}\cdot x+\boldsymbol{b}\cdot\boldsymbol{S}\cdot y+\boldsymbol{c}\cdot\boldsymbol{S}\cdot z=hk+ky+lz$ , therefore $\boldsymbol{F}(\boldsymbol{S})$ can be written as $\boldsymbol{F}(hkl)$ :

$$\boldsymbol{F}(hkl)=V\int_{x=0}^{1}\int_{y=0}^{1}\int_{z=0}^{1}\rho(xyz)\exp[2\pi i(hx+ky+lz)]dxdydz \ .$$

- The above equation expresses the structure factor as a function of the electron density. However, the aim of the crystallographer is to calculate the electron density of a molecule from the structure factors. The application of Fourier transformation solves this problem making it possible to calculate the electron density at any coordinate in the unit cell:

$$\rho(xyz)=\frac{1}{V}\sum_{h}\sum_{k}\sum_{l}\boldsymbol{F}(hkl)\exp[-2\pi i(hx+ky+lz)]$$

$$\rho(xyz)=\frac{1}{V}\sum_{h}\sum_{k}\sum_{l}|\boldsymbol{F}(hkl)|\exp[-2\pi i(hx+ky+lz)+i\alpha(hkl)]$$

- The term $|\boldsymbol{F}(hkl)|$ can be derived directly from the integrated experimental intensity of the individual reflections:

$I(int.,hkl)=K\times L\times P\times A\times|F(hkl)^2|$ , where $K$ implies various scale factors, $L$ is the Lorentz–factor, which is an instrument dependent factor, $P$ is a correction due to polarisation of the incident beam and $A$ is the correction for absorption.

Once the *amplitudes*, $|\boldsymbol{F}(hkl)|$ are obtained then the corresponding *phases* have to be determined. There are several methods to obtain phases for protein molecules of usual size. The most commonly used method is the multiple isomorphous replacement (MIR). The basic requirement of the method is to have one or several unique heavy atom binding sites in the molecule and isomorphism of the native and heavy atom soaked (derivative) crystals. The binding of a heavy atom introduces differences between the amplitudes of reflections collected from the native and the derivative crystals. On the basis of the isomorphous differences the bound metal ions can be located in the unit cell by calculating a difference Patterson function (see chapter 5.3). Phasing with the located

heavy atoms may provide sufficiently accurate starting phases to trace an initial electron density map.

With the advent of synchrotron radiation sources the multiwavelength anomalous dispersion (MAD) technique is becoming a more and more popular method to obtain initial phases. For light atoms like carbon, the reflections ($h\ k\ l$) and ($-h-k-l$) have the same intensity. However, for an atom with more electrons (heavy atoms) the Friedel–pairs are not equal anymore if the wavelength of the X–ray radiation is close to one of its absorption edges. The anomalous differences are usually small, and therefore precisely collected data of relatively high resolution are required. Once the anomalous scatterers are found the way of solving the structure is similar to MIR. Theoretically a single crystal is sufficient to measure all the necessary data if the anomalous scatterer is covalently bound in the molecule or incorporated into the crystal during growth.

Molecular replacement, a method which will be discussed later, can provide phases derived from a structurally similar protein molecule. Such a molecule has to possess high sequence, and as a consequence high structural homology to the protein of interest. In order to obtain a solution, the search molecule has to be properly placed into the unit cell of the target protein by a rotational and a subsequent translational search.

### 2.3.2 Data collection and processing

Crystals of sufficient size (> 150 μm) were transferred into a cryoprotecting solution which consisted of the original well solution plus 15% glycerol. The crystals seemed to be very stable during the solvent exchange thus no stepwise increase of the cryoprotectant was necessary. After ~5 minutes soak, the crystals were mounted in Hampton Research nylon loops and were flash frozen in a stream of dry nitrogen gas of 100 K (Cosier & Glaser, 1986). The freezing of crystals has several advantages (Hope,

1988; Henderson, 1990). Most importantly, the radiation damage is minimised compared to the case of a capillary mounted crystal. As a result in most cases a single crystal is sufficient to collect a whole data set at synchrotron sites. Another advantage is that the crystal does not have to be carried together with the whole crystallisation setup making the transport to the synchrotron site unproblematic. The frozen crystal has also higher mechanical stability compared to the capillary mounted crystal, therefore crystal slippage during data collection is not a problem anymore.

The data collection was carried out at DESY on beamline BW7A of the EMBL Outstation, Hamburg. A MAR Research 30 cm image plate was used to record the reflections. The data collection utilises the oscillation method (Arndt & Wonacott, 1977), which means that the crystal is rotated around an axis perpendicular to the X–ray beam, and the reflections passing through the Ewald–sphere are recorded. The disadvantage of this method is that reflections close to the spindle axis, depending on crystal symmetry and orientation, might never pass through the Ewald–sphere leaving a so called blind region, were the reflections are not recorded. A slight change in the crystal orientation can help to collect more complete data sets.

S1 nuclease crystals diffracted well to 1.7 Å resolution. Their diffraction properties are consistent with the primitive monoclinic space group $P2_1$, with cell dimensions $a$ = 42.1 Å, $b$ = 62.4 Å, $c$ = 101.3 Å and $\beta = 99.2°$. The cell volume to mass ratio (Matthews, 1968) suggested the presence of two molecules per asymmetric unit, resulting in a $V_M$ of 2.1 and a solvent content of 40%. Two passes of data collection were necessary because the exposure time required to collect good high resolution data caused overloading of the low resolution reflections. As a result a first data set from 1.7–50 Å was collected with an appropriately adjusted oscillation angle in order to avoid overlaps of reflections on the same image. The second low resolution pass from 5–50Å was collected with a uniform oscillation angle of 2 degrees.

The images were processed with the program MOSFLM (Leslie *et al.*, 1986). The program provides an X–windows based user interface (Campbell, 1995) besides the command line which provides an excellent way of interactively controlling the refinement process by the user. The resulting reflection files were scaled together by SCALA (Evans, 1997). The intensity values were converted to structure factors by TRUNCATE. The following table summarises the data collection statistics:

| ***Data set*** | ***Wavelength (Å)*** | ***Resolution range (Å)*** | ***Total number of reflections*** | ***Number of unique reflections*** | ***Overall $R_{sym}$ (%)*** | ***$R_{sym}$ in the highest resolution shell (1.70–1.79 Å) (%)*** |
|---|---|---|---|---|---|---|
| Native | 1.000 | 50–1.70 | 367047 | 56722 | 9.2 | 35.0 |
| ***Data set*** | ***Overall completeness (%)*** | ***Completeness in the highest resolution shell (1.70–1.79 Å) (%)*** | ***Overall I/σ*** | ***I/σ in the highest resolution shell (1.70–1.79 Å)*** | ***Mosaicity (°)*** | |
| Native | 99.2 | 99.4 | 6.9 | 2.2 | 0.5 | |

**Table 2.1** Data processing statistics for S1 nuclease. *I*, intensity. σ, standard deviation of the intensity. $R_{sym}=(\sum_{hkl}\sum_{i}|(I_{hkl}-\langle I\rangle_h)|)/\sum_{hkl}\sum_{i}|I_{hkl,i}|$ for *i* observations of a given reflection. <*I*>, mean intensity.

## 2.4 Molecular replacement

### 2.4.1 Introduction

Molecular replacement (MR) is the method to obtain phases when the atomic structure of a molecule with high structural homology to the protein of interest is

available. The model structure can be a crystal structure, but also can be derived from any modelling method yielding atomic coordinates. The model is placed in the unit cell of the target molecule by subsequent rotational and translational searches. In both cases an overlap function is calculated: the rotation function and the translation function respectively. The rotation function is the integral of the product of the Patterson–functions calculated from the model structure factors and the observed structure factors over a volume $U$ as formulated by Rossman & Blow (Rossmann & Blow, 1962):

$$R(\alpha,\ \beta,\ \gamma)=\int_U P(\boldsymbol{u})\times P_r(\boldsymbol{u}_r)\,d\boldsymbol{u}$$ , where α, β and γ are Eulerian angles. $P(\boldsymbol{u})$ = $P(u\ v\ w)$, where $u$, $v$, and $w$ are coordinates of the Patterson cell.

Substituting the Patterson–functions gives:

$$R(\alpha,\ \beta,\ \gamma)=\frac{1}{V^2}\sum_{\boldsymbol{h}}\sum_{\boldsymbol{h'}}|F(\boldsymbol{h})|^2|F[C](\boldsymbol{h'})|^2\times\int_U \exp[-2\pi i(\boldsymbol{h}+\boldsymbol{h'})\boldsymbol{u}]\,d\boldsymbol{u}$$ , where

$[C]$ is the rotation matrix bringing $\boldsymbol{u}$ to $\boldsymbol{u}_r$ and $\boldsymbol{h'}$ is the index of the reflections calculated from the search model in case of cross–rotation. The weighting term $\int_U \exp[-2\pi i(\boldsymbol{h}+\boldsymbol{h'})\boldsymbol{u}]\,d\boldsymbol{u}$ can be substituted by $\frac{U}{V}\times\boldsymbol{G}[-(\boldsymbol{h}+\boldsymbol{h'})]$ , which gives:

$$R(\alpha,\ \beta,\ \gamma)=\frac{U}{V^3}\sum_{\boldsymbol{h}}\sum_{\boldsymbol{h'}}|F(\boldsymbol{h})|^2|F[C](\boldsymbol{h'})|^2\times\boldsymbol{G}[-(\boldsymbol{h}+\boldsymbol{h'})]$$ , where $\boldsymbol{G}$ is the Fourier–transform of a sphere of volume $U$. $\boldsymbol{G}$ is a function which falls very rapidly for values of $\boldsymbol{h'}$ differing from $-\boldsymbol{h}$, considerably decreasing the number of terms to be calculated. Crowther (Crowther, 1972) formulated the fast rotation function by expanding the Patterson functions in terms of spherical harmonics instead of Cartesian Fourier components. This formulation resulted in a hundred–fold improvement in computational speed compared to the original procedure of Rossmann and Blow.

Once the solutions from the rotational search have been determined, a translational search is carried out. This can be done in a trial–and–error procedure moving the search

molecule in the unit cell and calculating an *R*–factor or a correlation coefficient as a function of the molecular position. In another method described by Crowther and Blow (Crowther & Blow, 1967) a translation function is calculated that gives the correlation between a set of cross–Patterson vectors for a model structure and the observed Patterson–function:

$$T(\boldsymbol{t})=\int_V P_{1,2}(\boldsymbol{u},\boldsymbol{t})\times P(\boldsymbol{u})\,d\boldsymbol{u}$$ , where $P(\boldsymbol{u})$ is the observed Patterson–function, and $P_{1,2}(\boldsymbol{u},\boldsymbol{t})$ is the cross–Patterson–function of the model structure in which two molecules are related by crystallographic symmetry. With the expansion of the Patterson functions the following formula of the translation function can be derived:

$$T(\boldsymbol{t})=\sum_{\boldsymbol{h}} \left|F_{obs}(\boldsymbol{h})\right|^2\cdot \boldsymbol{F}_M(\boldsymbol{h})\cdot \boldsymbol{F}_M^*(\boldsymbol{h}\cdot[C])\exp[-2\pi i\boldsymbol{h}\cdot\boldsymbol{t}]$$ , where $\boldsymbol{F}_M$ is the structure factor of the model molecule, $\boldsymbol{F}_M$* is its complex conjugate, [*C*] is the rotation matrix of crystallographic symmetry, and $\boldsymbol{t}$ is the translation. The translation function can be corrected for the unwanted self–Patterson vectors:

$$T_1(\boldsymbol{t})=\sum_{\boldsymbol{h}} \{\left|F_{obs}(\boldsymbol{h})\right|^2-\sum_{n=1}^{n}\left|F_{M(n)}(\boldsymbol{h})\right|^2\}\cdot \boldsymbol{F}_M(\boldsymbol{h})\cdot \boldsymbol{F}_M^*(\boldsymbol{h}\cdot[C])\exp[-2\pi i\boldsymbol{h}\cdot\boldsymbol{t}]$$ , where $n$ is the number of molecules in the cell. The functions $T(\boldsymbol{t})$ and $T_1(\boldsymbol{t})$ are the product functions of two correctly oriented molecules in the unit cell. Taking all the possible intermolecular vectors into account the three dimensional expression of the translation function can be derived:

$$T_2(\boldsymbol{m})=\sum_{\boldsymbol{h}} \left|F_{obs}(\boldsymbol{h})\right|^2\sum_{j=1}^{n}\sum_{k=1}^{n}\left|F_M(\boldsymbol{h}[C_j])\right|^2\cdot \boldsymbol{F}_M^*(\boldsymbol{h}[C_k])\times\exp[-2\pi i\boldsymbol{h}(\boldsymbol{d}_j-\boldsymbol{d}_k)]\ \times$$

$$\times\ \exp[-2\pi i\boldsymbol{h}([C_j]-[C_k])\boldsymbol{m}]\ .$$

The signal to noise ratio can be further improved by subtracting the self–Patterson vectors. The application of negative B–factors to the structure factors gives further improvement by sharpening the Patterson–map.

Once the correct solutions are found, an electron density map is calculated using

the measured structure factors and the phases calculated from the model. The usage of the model phases implies that the electron density map is fairly much biased towards the model's electron density. Therefore, care has to be taken to get rid of this bias during the refinement.

## 2.4.2 Application to S1 nuclease

As it was already mentioned, S1 nuclease possesses high sequence homology to P1 nuclease from *P. citrinum,* so it was straightforward to try molecular replacement with the refined P1 nuclease structure as a model. The similarity in function of the two nucleases is also indicative of high structural homology. The P1 structure deposited as the entry 1AK0 in the PDB (Romier *et al.*, 1998) was used as the search model. Only the peptide atoms were included in the calculation, all the water molecules, the carbohydrate side chains and even the zinc ions were removed. The CCP4 program AMoRE (Navazza, 1994) was used for the calculation of the cross–rotation and translation function including reflections between 4 and 20 Å with a properly chosen Patterson radius. The first ten solutions were accepted to calculate the translation function. Since it was known that there were two molecules per asymmetric unit therefore another translation search was run fixing one of the model positions corresponding to the highest correlation and lowest *R*–factor. Rigid body fitting resulted in a correlation coefficient of 51.7 and an *R*–factor of 44.4%. The search model coordinates were transformed with respect to the two sets of rotational and translational parameters resulting in two correctly placed molecules in the asymmetric unit.

## 2.5 Density modification

### 2.5.1 Introduction

Frequently the initial electron density map calculated with the derivative or model phases are not or hardly interpretable. Prior to model building and refinement density modification methods can be applied to improve the quality of the electron density map making the interpretation easier (Podjarny, 1985; Podjarny *et al.*, 1987). Density modification methods are aimed at improving the agreement between the electron density calculated from experimentally derived structure factors and a set of physical constraints based on known characteristics of the density function. During density modification all available structural information should be used (Brünger & Nilges, 1993). In the following the methods utilising a particular set of structural information will be described.

**Solvent flattening**

It is known from highly refined structures that the solvent region of the electron density map is rather flat, and has a low density value due to the dynamic nature of the solvent molecules, which results in a time–averaged electron density. If the region occupied by the protein is identified, the electron density of the solvent can be set to the theoretical average value. As a result the noise is reduced in the density map in general. The identification of protein region in the electron density map can be done manually by defining a mask, i.e. a molecular envelop around the protein, which is not always easy in case of a noisy map. To address the problem of subjectivity an automated method was proposed by Wang (Wang, 1985) and modified by Leslie (Leslie, 1987). In the Wang–method a grid is superimposed on the cell. At each grid point the density is replaced by a new density value that is proportional to the weighted sum of densities within a sphere of radius $R$ centred at the grid point. In the summation, density less then zero has a weight of

zero, while density higher than zero has a weight of $1-r_{ij}/R$, where $r_{ij}$ is the grid spacing. In the subsequent iteration steps the molecular envelope and the Wang–radius $R$ are updated, and, in case one is using experimental phases, higher resolution reflections are included in the calculation (phase extension).

**Solvent flipping**

This method is similar to solvent flattening, but the solvent density values are not set simply to an average value, rather the inverted solvent density values are added to the initial map. This is similar to adding negative noise to an image in order to strengthen the signal/noise ratio (Abrahams & Leslie, 1996).

**NCS–averaging**

The molecules in the asymmetric unit related by non–crystallographic symmetry has basically the same electron density except some local variations due to different molecular contacts in the crystal. The quasi equal density imposes a constraint on the structure factors, and as a consequence on the phase angles, which can be used to calculate better quality maps. Bricogne (Bricogne, 1974) developed a successful algorithm to carry out NCS–averaging. The first step in the procedure is the definition of a molecular envelop around one of the monomers. The mask is then replicated around the NCS–related molecules applying the already known NCS operators. Then the electron density in the asymmetric unit is averaged, the solvent region is flattened and the asymmetric unit is reconstituted. The resulting map is back–transformed giving new calculated phases which are combined with the starting phases. A new map is calculated using experimental structure factors and the combined phases. The procedure is then repeated from the definition of a new molecular envelope.

**Histogram matching**

Histogram matching is a method originally applied in image processing to improve

an image by modifying the histogram of densities for that image to match the histogram expected from a perfect image. Histogram matching is usually applied together with solvent flattening. In case of proteins, the distribution of electron density values (the histogram) appears to be fairly independent from the nature of the protein at the same resolution. The frequency distribution of density levels of a high resolution density map can be used to modify the histogram of a poorer quality map (Zhang & Main, 1988). The electron density values in the maps are histogrammed into a number of equally spaced bins. A scale factor and a shift parameter is then applied to the probability distribution of the map to be modified in order to match the histogram of the high quality map.

### 2.5.2 Application to S1 nuclease

The CCP4 program SFALL was used to calculate structure factors and phases using the initial model structure from molecular replacement. Prior to density modification the CCP4 program SIGMAA (Read, 1986) was used to calculate weighted Fourier coefficients in order to reduce the model bias. The first electron density map calculated this way was fairly well interpretable, but further improvements could be achieved by the application of the CCP4 program DM (Cowtan, 1994).

In the DM calculations all reflections were included. A mask was calculated around one of the model monomers to do NCS averaging. In addition to NCS averaging, solvent flattening and histogram matching were applied for 50 cycles of density modification. The refined NCS–operator matrix was used to run another fifty cycles of calculation. The difference between the model phased but $\sigma_A$ weighted electron density map and the map after density modification and NCS averaging is illustrated in Figure 2.6.

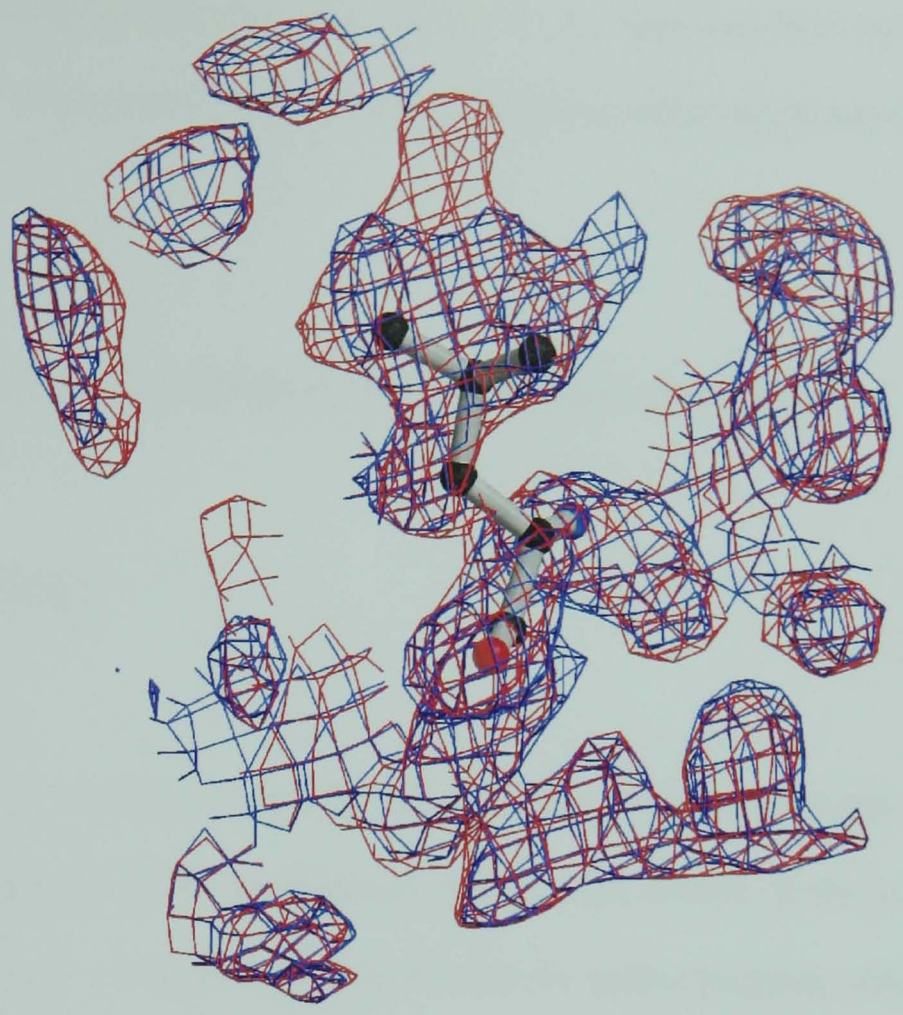


**Figure 2.6** Electron density map calculated before (blue) and after (red) density modification around residue 59. Residue 59 is a leucine (as shown) in the model structure, whereas it is a tyrosine in S1 nuclease. The blue map is calculated with $\sigma_A$–weighted Fourier coefficients and model phases (SIGMAA), while the red map is calculated with combined phases from density modification (DM). The map after DM (red) shows clearly the shape of a tyrosyl side chain.

On the basis of the map output by DM the side chains of the model molecule were substituted with the S1 nuclease sequence utilising an automated procedure of the program Xfit (McRee, 1999). The program not only mutates the sequence, but tries to do real–space fitting of the side chains with quite good success in the case of S1. The mutation I120T mentioned in the paper on the cloning of S1 was identified (Lee *et al.*, 1995). Only a short loop region between residue 98 and 106 and the last three residues (265, 266 and 267) were not visible in the electron density map. Six zinc ions corresponding to peaks at 6 $\sigma$

level were also immediately identified in the $F_o-F_c$ map and built into the model. The second molecule in the asymmetric unit was reconstructed by applying the NCS symmetry operation.

## 2.6 Refinement and validation of S1 nuclease structure

### 2.6.1 Introduction

The initial molecular model, even if all amino acids are included, always has errors in the atomic coordinates and temperature factors. The source of the coordinate errors is the quality of the electron density map itself used for model building, due to lack of atomic resolution and inaccuracy in phases, and the atomic B–factors which are almost always incorrectly set prior to initial building. In order to reach the best possible correlation between the observed and calculated structure factors, refinement of atomic coordinates and B–factors is necessary.

#### 2.6.1.1 Observations vs. refined parameters

The quantity of the available experimental data strongly influences the choice of refinement strategy. Having collected data to ultra high resolution ensures *unrestrained* refinement of the structure of interest. However, in most of the cases the quantity of experimental data (the number of reflections) is relatively low compared to the parameters to be refined. To improve the ratio between experimental data and parameters, stereochemical restraints can be applied to the atomic coordinates in the form of ideal bond length, bond angles, torsion angles, etc. derived from high resolution small molecule X–

ray structures. If NCS is present, NCS–restraints can be applied to the coordinates and temperature factors. The refinement of temperature factors is also highly dependent on the quantity of experimental data, in other words on the data resolution. With increasing resolution of the data the number of B–factors refined per residue can be gradually increased. As a rule of thumb an overall B–factor is assigned to all atoms in a residue if the data have less than 3 Å resolution. Between 3 and 2.5 Å resolution group–based B–factors are calculated: one B–factor for all main chain atoms and another for all side chain atoms. Higher resolution data make the refinement of individual isotropic B–factors possible, although it might be necessary to apply restraints. If the data resolution is better than 1.2 Å, individual anisotropic B–factors can be refined, which take the non–isotropic thermal motion of atoms into account. On the other hand it is also important to maximise the number of reflections involved in the refinement by using the low order reflections too. Earlier the low order reflections, being seriously affected by the disordered solvent region, were usually omitted from the refinement by simply applying a low resolution cutoff. With the proper correction for the bulk solvent the low order reflections can be and must be included in refinement.

### 2.6.1.2 Conventional refinement

The aim of refinement is the minimisation of the total potential energy, which, in general, consists of an empirical and an X–ray term:

$$E_{total} = E_{xray} + E_{emp}.$$

In the case when NCS is present and NCS–restraints are applied, an energy term $E_{NCS}$ has to be introduced. Minimisation of $E_{xray}$ implies the fitting of calculated structure factors to the observed ones through the optimisation of atomic coordinates and temperature factors. $E_{emp}$, depending on the implementation, consists of weighted energy terms related to the

applied stereochemical restraints. If the observation–parameter ratio is sufficiently high, an *unrestrained* refinement can be carried out. In such case the empirical term is not calculated, only the X–ray term is minimised.

In most of the current refinement programs two different approaches for the minimisation of $E_{xray}$ are implemented. The traditional refinement uses the method of least squares, which is the minimisation of the following function: $\sum_{hkl} w(hkl)(|F_o(hkl)|-|F_c(hkl)|)^2$ . The goal of the refinement is to optimise the atomic parameters resulting in $F_c$ as close as possible to $F_{obs}$.

In the last few years the so called maximum likelihood refinement (MLR) has been proven to be a useful and powerful method for X–ray structure refinement (Bricogne & Irwin, 1996; Read, 1996; Pannu & Read, 1996; Murshudov *et al.*, 1997; Pannu *et al.*, 1998). The main difference between MLR and the least squares refinement lies in the goal of the refinement. In contrast to least squares refinement, MLR tries to maximise the chances (the probability) to improve the model structure further, instead of fitting the calculated structure factors to the observed ones. Other advantages of MLR are how it handles the experimental errors in the observed magnitudes, and the partiality of the model structure.

### 2.6.1.3 Refinement using molecular dynamics

Conventional refinement based on either the least–squares or maximum likelihood method minimises the total potential energy until a local minimum is found. The local minimum, especially when refining an initial model, does not necessarily coincide with the global one. The refinement method which allows potential energy barriers to be crossed by moving uphill on a potential energy surface is molecular dynamics (MD) (Brünger *et al.*,

1987; Brünger & Nilges, 1993). Molecular dynamics simulates the movement of atoms in the molecule over specified time intervals at a certain temperature by solving Newton's equation of motion:

$$m_i\left(\frac{d^2\boldsymbol{r}_i}{dt^2}\right)=\frac{\partial E_{total}}{\partial \boldsymbol{r}_i} \quad .$$

In the case of crystallographic refinement the energy term $E_{total}$ includes the X–ray energy term as well. Molecular dynamics is usually applied as simulated annealing (SA) which allows an extensive exploration of the multiparameter target function, $E_{total}$, helping the global minimum to be located (Brünger, 1988; Brünger *et al.*, 1990, Brünger *et al.*, 1997). In a simulated annealing calculation the starting temperature, thus the kinetic energy of the atoms, is very high allowing large energy barriers to be crossed. The temperature of the system is then slowly cooled down (annealed). Slow cooling helps to ensure that the global, and not a local minimum, of potential energy is found. Simulated annealing has a larger radius of convergence than conventional refinement. Therefore its application is very helpful as one of the first steps in the refinement process. Crystallographic MD refinement is implemented in several computer programs, like X–PLOR (Brünger, 1992), CNS (Brünger *et al.*, 1998) and GROMOS (Fujinaga *et al.*, 1989).

### 2.6.1.4 Monitoring the progress of refinement

The correlation between $F_o$ and $F_c$ can be monitored by the conventional crystallographic $R$–factor:

$$R=\frac{\sum_{hkl}\left|F_o(hkl)-w\cdot F_c(hkl)\right|}{\sum_{hkl}F_o(hkl)} \quad \text{, where} \quad w=\frac{\sum_{hkl}F_o(hkl)}{\sum_{hkl}F_c(hkl)} \quad .$$

A decrease of $R$ indicates a better correlation between observed and calculated structure factors, but it does not necessarily indicate the correctness of the refinement. It has been

shown that an incorrect model can also be refined to a fairly good $R$ values (Brändén & Jones, 1990). As a result the $R_{free}$ concept has been introduced by Brünger (Brünger, 1992). $R_{free}$ is calculated analogously to $R$ from a *test set* of reflections which are not used for the refinement. Reflections used in the refinement are termed as the *working set* of reflections. As it has been shown, $R_{free}$ strongly correlates with the errors in the model phases, therefore its usage together with $R_{work}$ is a better indicator of the correctness of the model than just the $R_{work}$ alone (Brünger, 1992). Practically it means, if the $R_{work}$ is decreasing but the $R_{free}$ settles or increases, the model is *overrefined.* The $R_{free}$ values can also be used to optimise refinement parameters, like weight factors (Kleywegt & Brünger, 1996). It is important to emphasise that besides monitoring $R_{work}$ and $R_{free}$ as quality indicators, one has to check also the deviations from ideal stereochemistry in the model structure.

### 2.6.2 The refinement of S1 nuclease structure

As described in a previous paragraph, after several cycles of density modification, the sequence of P1 nuclease, used as a search model in MR, was modified to match the sequence of S1 nuclease. The excellent quality of the electron density map allowed the real–space fitting and easy manual correction of the side chain conformations.

In the first part of the refinement process only the CNS program suite was used (Brünger *et al.*, 1998). Since the protein model was almost complete, all reflections were included in the subsequent calculations. Initially medium restraints were imposed between the NCS–related monomers. The NCS–restraints were gradually decreased during refinement. The first refinement step was rigid body refinement defining the two monomers in the asymmetric unit as rigid bodies. The rigid body refinement was followed by SA utilising torsion angle molecular dynamics (Rice & Brünger, 1994) and a maximum likelihood target (Adams *et al.*, 1997). The model was heated to 2500 K and cooled in

steps of 25 K to a final temperature of 300 K. In the following step a two B–factor per residue, group–based B–factor refinement was run, followed by the calculation of NCS–averaged $\sigma_A$–weighted electron density maps $2F_o–F_c$ and $F_o–F_c$ (Read, 1986). The resulting electron density maps allowed the region between residue 98 and 106 to be rebuilt and allowed residue 264 to the C–termini of the two chains to be added using the program Xfit (McRee, 1999). The manually modified model was subjected to 100 cycles of positional refinement, followed by 50 cycles of restrained isotropic individual B–factor refinement and electron density map calculation. The first hundred water molecules corresponding to the highest peaks of the $F_o–F_c$ map and six *N*–acetyl–glucosamine residues were manually added using the program Xfit. This program searches for water positions automatically providing a quick and easy way of inspecting and occasionally deleting incorrectly positioned waters. Xfit does not place water molecules in the proximity of metal ions, therefore the otherwise extremely well defined water ligands of the zinc ions were placed manually. The PDB files of NAG residues were downloaded from the HIC–Up database (http://alpha2.bmc.uu.se/hicup/). Three β–C1–O4–linked pairs of NAG residues could be easily built into the difference density which was nicely continuous from the ND2 atoms of N92 and N228 with the exception of N228 in chain A. The missing density at N228A is due to the different molecular environment at the glycosylation sites of the two S1 monomers. Several cycles of rebuilding and manual addition of waters followed by positional and isotropic individual B–factor refinement were performed. As soon as ~200 water molecules were built into the model the refinement was carried on with the combination of CNS and the CCP4 program REFMAC, now without NCS–restraints (Murshudov *et al.*, 1997). REFMAC refines geometry and temperature factors simultaneously by using a maximum likelihood residual. CNS was used to calculate partial structure factors as the contribution from the solvent. Before each model inspection step REFMAC was run for 15 cycles running PROTIN

(Hendrickson, 1985) after every three internal cycles. Since the two programs use incompatible file formats, several UNIX shell scripts had to be written to combine all the data in a single MTZ file prior to REFMAC runs.

After several cycles by REFMAC using the mask–based bulk solvent correction of CNS the refinement and model building had converged with an $R_{work}$ of 16.3% and $R_{free}$ of 19.6%. The final model contains two protein chains in the asymmetric unit (Figure 2.7). Chain A has 264 residues, while in chain B one additional C–terminal residue could be identified. There are six zinc ions bound per monomer; three in the active centre and three others are involved in crystal contacts, coordinated from neighbouring S1 molecules. Three carbohydrate side chains could be built into the model. In chain A there are only two NAG residues bound to N92, whereas in chain B in addition to the two NAG residues there are three mannose residues. The content of the present model and some refinement statistics are shown in Table 2.2.

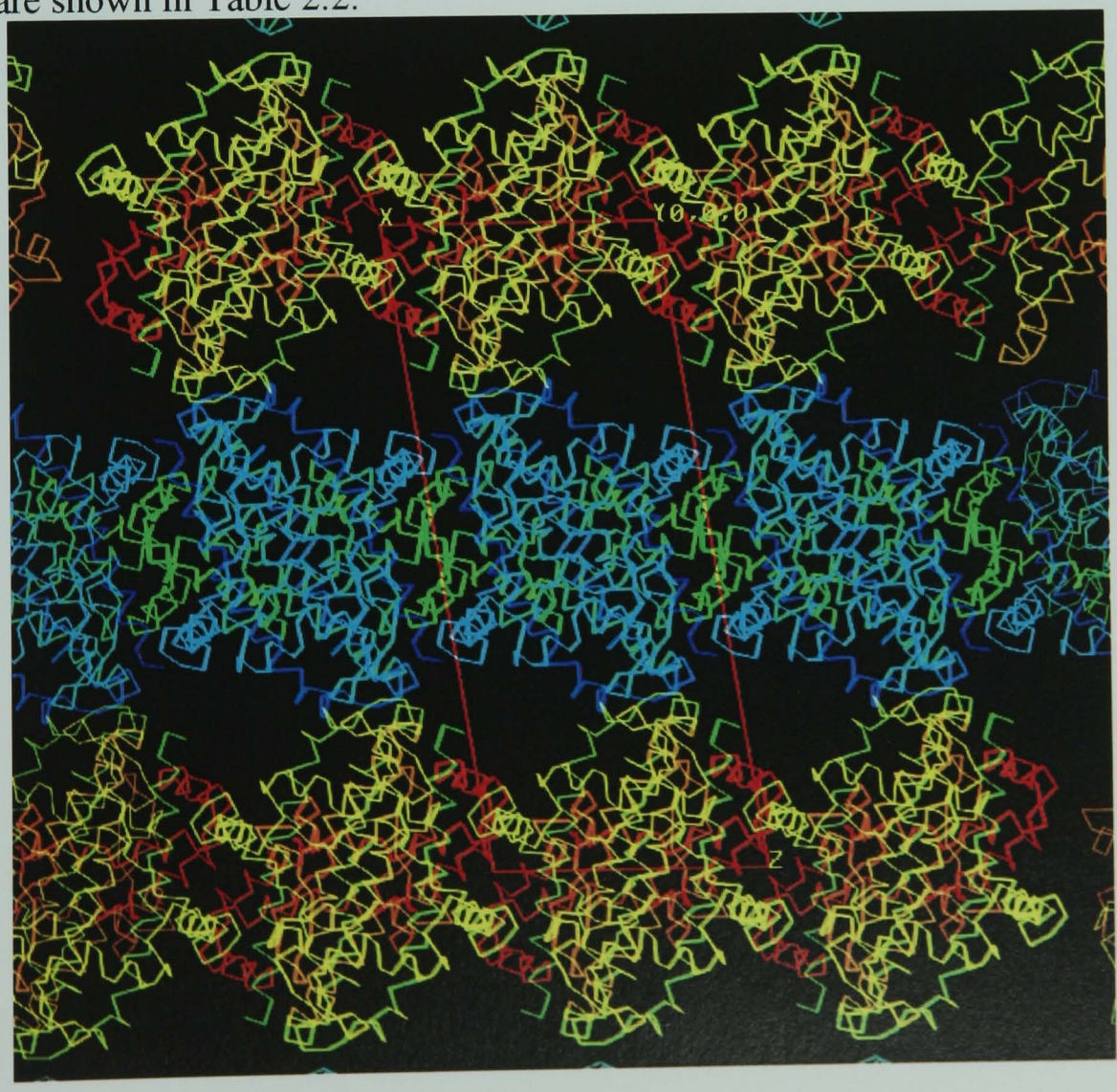



**Figure 2.7** The packing of S1 molecules in the crystal lattice. The orientation is chosen as

the two–fold axis is perpendicular to the picture's plane. The monomers of the non–crystallographic dimers can be distinguished by colouring.

| **Refinement statistics** | |
|---|---|
| Total number of reflections used | 57211 |
| Working set of reflections | 53824 |
| R–factor (%) | 16.34 |
| Test set of reflections | 2877 |
| R–free (%) | 19.68 |
| Total number of protein atoms | 4063 |
| Total number of carbohydrate atoms | 117 |
| Total number of zinc atoms | 12 |
| Total number of water molecules | 582 |
| **Geometry statistics** | |
| R.m.s.Δ bond distance (Å) | 0.009 |
| R.m.s.Δ bond angle (Å) | 0.02 |
| B–factor R.m.s.Δ ($Å^2$) | |
| Bonded main chain atoms ($Å^2$) | 1.09 |
| Bonded side chain atoms ($Å^2$) | 1.44 |
| Angle main chain atoms ($Å^2$) | 2.31 |
| Angle side chain atoms ($Å^2$) | 3.09 |
| Average B factor ($Å^2$) | |
| Main chain atoms ($Å^2$) (A, B) | 9.31 (9.70, 8.91) |
| Side chain atoms ($Å^2$) (A, B) | 10.47 (10.73, 10.21) |
| All protein atoms ($Å^2$) (A, B) | 10.28 (10.64, 9.92) |
| Zinc atoms ($Å^2$) (A, B) | 8.85 (9.07, 8.63) |
| Carbohydrate atoms ($Å^2$) | 23.67 |
| Water molecules ($Å^2$) | 21.07 |
| **Non–crystallographic symmetry** | |
| R.m.s.Δ $C_\alpha$ (Å) | 0.209 |
| B–factor R.m.s.Δ of all atoms ($Å^2$) | 2.319 |

**Table 2.2** Refinement and geometry statistics of the S1 nuclease model.

## 2.6.3 Validation of the refined structure

The electron density map using the refined model was of good quality, without unexplained positive or negative difference density features. The quality of the model was analysed by the structure validation programs PROCHECK (Laskowski *et al.*, 1993) and WHATIF (Vriend, 1990). WHATIF suggested some minor corrections to the model like flipping amino acid side chains to obey torsion angle conventions and removal of a few water molecules. The Ramachandran plot calculated by PROCHECK is shown on Figure 2.8.

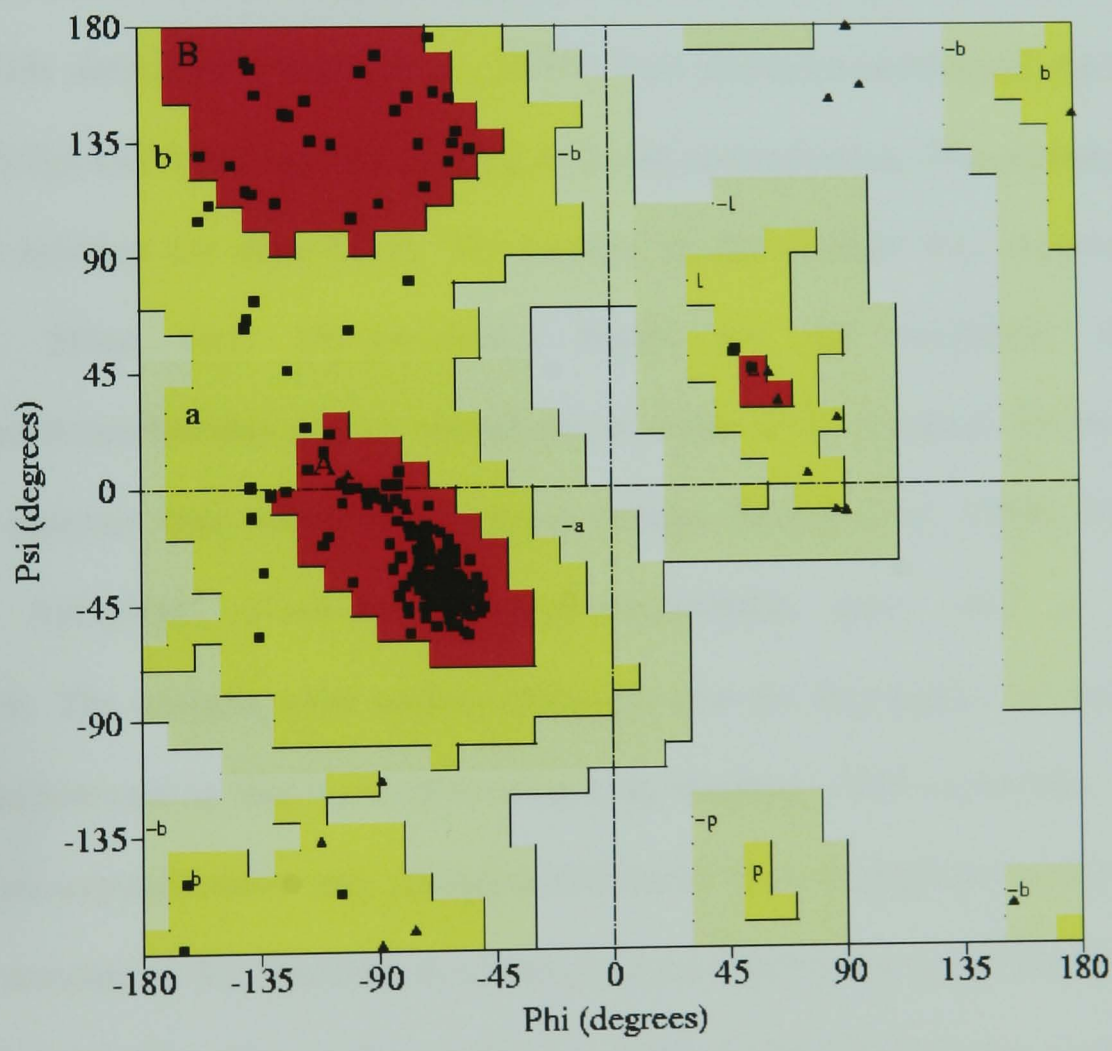



**Figure 2.8** Ramachandran plot for chain B of the refined model of S1 nuclease. There are no residues in the disallowed region of the plot (white). Most of the residues have phi–psi values in the most favoured region (red), the rest are in the allowed regions (bright yellow). Glycine residues are represented by black triangles, the phi–psi value of other residues are shown as black squares.

## 2.7 Substrate binding studies

One of the major goals of this work was to obtain complex structures of S1 nuclease with uncleavable substrate analogues. Soaking and co–crystallisation with such substrate analogues were carried out to produce complex crystals. Crystals of S1 were produced as described in chapter 2.2.3.2. Two types of substrate analogues, phosphorodithioates and 2'–*O*–methyloligoribonucleotides, were used for the soaking experiments. In phosphorodithioates the two oxygen atoms which are not involved in the ester bond formation are exchanged to sulphur, resulting in a completely nuclease resistant oligonucleotide derivative (Eldrup *et al.*, 1994). Two phosphorodithioates, $Ap(S)_2T$ and $Ap(S)_2Tp(S)_2Tp(S)_2T$ were used for soaking at 2 mM concentration. The crystals were left in the soak solution for days (1–7). No damage of the crystals was observed due to soaking. Since their phosphodiester bonds are not modified, the 2'–*O*–methyloligoribonucleotides are not completely resistant to S1 nuclease but they are still much more resistant than unmodified oligonucleotides (Sproat *et al.*, 1989). For soaking ApU and ApUpUpU 2'–*O*–methyloligoribonucleotides were used at 10 mM concentration. The crystals were soaked overnight and the degradation of the substrate analogue was assessed by thin layer chromatography showing ~30% conversion.

For co–crystallisation only phosphorodithioates were considered because they are completely resistant to S1, therefore no cleavage occurs during the long time required for crystal formation. Five fold molar excess of $Ap(S)_2Tp(S)_2Tp(S)_2T$ was used for co–crystallisation using the same crystallisation condition described in chapter 2.2.3.2.

Crystals from both soaking and co–crystallisation were flash frozen in liquid nitrogen using 15% glycerol as cryoprotectant. Complete data up to 2.3 Å resolution were collected from frozen (100 K) derivative crystals using a rotating anode as X–ray source. The data sets were processed with MOSFLM (Leslie *et al.*, 1986). The derivative crystals

had essentially identical cell parameters compared to the high resolution native data. Difference Fourier maps with $F_o$(derivative)–$F_o$(native) as coefficients were calculated using model phases of the refined S1 nuclease model. Prior to map calculations the derivative data were scaled to the high resolution native with SCALEIT (CCP4, 1994). The inspection of the electron density maps, however, did not reveal any new density features corresponding to a bound oligonucleotide derivative. Further refinement of the S1 nuclease structure with REFMAC against these "derivative" data sets and the calculation of clearer $2F_o-F_c$ and $F_o-F_c$ maps did not change the situation. These results indicate that either the binding constants of such complexes are very weak or the binding is unfavoured at the condition of crystallisation. Unfortunately, crystals do not form at the pH of highest activity (4.6) where binding is expected to be more favoured.

## 2.8 References

Abrahams, J.P. & Leslie, A.G.W. (1996). Methods used in the structure determination of bovine mitochondrial $F_1$ ATPase. *Acta Crystallogr.* **D52**, 30–42.

Arndt, U.W. & Wonacott, A.J. (1977). *The Rotation Method in Crystallography.* North–Holland publishing company, Amsterdam

Blundell, T.L. & Johnson, L.N. (1976). *Protein crystallography.* Academic Press, London

Brändén, C.I. & Jones, A. (1990). Between objectivity and subjectivity. *Nature* **343**, 687–689.

Bricogne, G. (1974). Geometric sources of redundancy in intensity data and their use for phase determination. *Acta Crystallogr.* **A30**, 395–405.

Bricogne, G. & Irwin, J. (1996). *Macromolecular refinement: Proceedings of the CCP4 Study Weekend*, Dodson, E., Moore, M, Ralph, A. & Bailey, eds., pp. 85–92. Daresbury Laboratory, Warrington, UK.

Brünger, A.T., Kuriyan, J. & Karplus, M. (1987). Crystallographic R–factor refinement by molecular dynamics. *Science* **235**, 458–460.

Brünger, A.T., Krukowski, A. & Erickson, J. (1990). Slow–cooling protocols for crystallographic refinement by simulated annealing. *Acta Crystallogr.* **A46**, 585–593.

Brünger, A.T. (1992). *X–PLOR Version 3.1.* Yale University, New Haven.

Brünger, A.T. (1992). Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472–475.

Brünger, A.T. & Nilges, M. (1993). Computational challenges for macromolecular structure determination by X–ray crystallography and solution NMR–spectroscopy. *Quarterly Rev. Biophys.* **26**, 49–125.

Brünger, A.T., Adams, P.D. & Rice, L.M. (1997). New applications of simulated annealing in X–ray crystallography and NMR. *Structure* **5**, 325–336.

Brünger, A.T. et al. & Warren, G.L. (1998). Crystallography & NMR System: a new software suit for macromolecular structure determination. *Acta Crystallogr.* **D54**, 905–921.

Campbell, J.W. (1995). XDL_VIEW, an X–windows–based toolkit for crystallographic and other applications. *J. Appl. Cryst.* **28**, 236–242.

Carter, C.W.Jr. & Carter, C.W. (1979). Protein crystallisation using incomplete factorial experiments. *J. Biol. Chem.* **254**, 12219–12223.

Carter, C.W.Jr. (1992). Design of crystallisation experiments and protocols. In *Crystallisation of nucleic acids and proteins. A practical approach.* Ducroix, A. & Giege, R., Eds., Oxford University Press, New York, pp. 47–69.

Collaborative Computer Project No. 4 (1994). The CCP4 Suite: Programs for Protein Crystallography. *Acta Crystallogr.* **D50**, 760–763

Cosier, J. & Glazer, A.M. (1986). A nitrogen gas stream cryostat for general X–ray diffraction studies. *J. Appl. Cryst.* **19**, 105–107.

Cowtan, K. (1994). Dm: An automated procedure for phase improvement by density modification. *Joint CCP4 and ESF–EACBM Newsletter on Protein Crystallography* **31**, 34–38.

Crowther, R.A. (1972). In *The Molecular Replacement Method*, pp. 173–178; Rossmann, M.G., ed. Gordon & Breach, New York

Crowther, R.A. & Blow, D.M. (1967). Amethod of positioning a known molecule in an unknown crystal structure. *Acta Crystallogr.* **23**, 544–548.

Ducroix, A. & Giege, R. (1992). *Crystallisation of nucleic acids and proteins. A practical approach.* Oxford University Press, New York

Evans, P.R. (1997). Scala. *Joint CCP4 and ESF–EACBM Newsletter* **33**, 22–24.

Fehér, G. & Kam, Z. (1985). Nucleation and groth of protein crystals: General principles and assays. *Meth. Enzymol.* **114**, 77–112.

Fujinaga, M., Gros, P. & Van Gunsteren, W.F. (1989). Testing the method of crystallographic refinement using molecular dynamics. *J. Appl. Crystallogr.* **22**, 1–8.

Gill S.C. & von Hippel P.H. (1989). Calculation of protein extinction coefficients from amino acid sequence data. *Anal. Biochem.* **182**, 319–26.

Gruner, S.M. & Ealick, S.E. (1995). Charge coupled device X–ray detrectors for macromolecular crystallography. *Structure* **3**, 13–15.

Henderson, R. (1990). Cryo–protection of protein crystals against radiation damage in electron and X–ray diffraction. *Proc. Royal Soc. Lond.* **B241**, 6–8.

Hendrickson, W.A. (1985). Stereochemically restrained refinement of macromolecular structures. *Meth. Enzymol.* **115**, 252–270.

Hope, H. (1988). Cryocrystallography of biological macromolecules: a generally applicable method. *Acta Crysallogr.* **44**, 22–26.

Jancarik, J. & Kim, S.H. (1991). Sparse matrix sampling: a screening method for crystallisation of proteins. *J. Appl. Cryst.* **24**, 409–411.

Kleywegt, G.J. & Brünger, A.T. (1996). Checking your imagination: applications of the free R value. *Structure* **4**, 897–904.

Laskowski, R.A., MacArthur, M.W., Moss, D.S. & Thornton, J.M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26**, 283–291.

Lee, B.R, Kitamoto, K., Yamada, O. & Kumagai, C. (1995). Cloning, characterization and overproduction of nuclease S1 gene (*nucS*) from *Aspergillus oryzae*. *Appl. Microbiol. Biotechnol.* **44**, 425–431.

Leslie, A.G.W., Brick, P. & Wonacott, A.T. (1986). MOSFLM. *Daresbury Lab. Inf. Quart. Protein. Cryst.* **18**, 33–39.

Leslie, A.G.W. (1987). A reciprocal–space method for calculating a molecular envelope using the algorithm of B.C. Wang. *Acta Crystallogr.* **A43**, 134–136.

McPherson, A. (1982). *Preparation and analysis of protein crystals.* John Wiley, New York

McPherson, A. (1985). Use of polyethylene glycol in crystallisation of macromolecules. *Meth. Enzymol.* **114**, 120–125.

McRee, D.E. (1999). XtalView/Xfit – A versatile program for manipulating atomic coordinates and electron density. *J. Struct. Biol.* **125**, 156–165.

Matthews, B.W. (1968). Solvent content of protein crystals. *J. Mol. Biol.* **33**, 491–497.

Murshudov, G.N., Vagin, A.A. & Dodson, E. (1997). Refinement of macromolecular structures by the Maximum–Likelihood Method. *Acta Crystallogr.* **D53**, 240–255.

Navaza, J. (1994). AMoRE: an automated package for molecular replacement. *Acta Crystallogr.* **A50**, 157–163.

Oldfield, T.J., Ceska, T.A. & Brady, R.L. (1991). A flexible approach to automated protein crystallisation. *J. Appl. Cryst.* **24**, 255–260.

Pannu, N.S. & Read, R.J. (1996). Improved structure refinement through maximum likelihood. *Acta Crystallogr.* **A52**, 659–668.

Pannu, N.S., Murshudov, G.N., Dodson, E.J. & Read, R.J. (1998). Incorporation of prior phase information strengthens maximum likelihood structural refinement. *Acta Crystallogr.* **D54**, 1285–1294.

Podjarny, A.D. (1985). Density modification methods. In *Crystallography in Molecular Biology*. Series A, **126**, Plenum Press

Podjarny, A.D., Bhat, T.N. & Zwick, M. (1987). Improving crystallographic macromolecular images: the real–space approach. *Ann. Rev. Biophys. Biophys. Chem.* **16**, 351–373.

Read, R.J. (1986). Improved Fourier coefficient s for maps using phases from partial structures with errors. *Acta Crystallogr.* **A42**, 140–149.

Read, R.J. (1997). Model Phases: probabilities and bias. *Meth. Enzymol.* **277**, 110–128.

Rice, L.M. & Brünger, A.T. (1994). Torsion angle dynamics: reduced variable conformational sampling enhances crystallographic structure refinement. PROTEINS: *Structure, Function and Genetics* ***19***, 277–290.

Romier, C. et al. & Suck, D. (1998). Recognition of single–stranded DNA by nuclease P1: high resolution crystal structure of complexes with substrate analogs. *PROTEINS: Structure, Function and Genetics* **32**, 414–424.

Rossmann, M.G. & Blow, D.M. (1962). The detection of sub–units within the crystallographic asymmetric unit. *Acta Crystallogr.* **15**, 24–31.

Tarentino, A.L. & Maley, F. (1974). The release of intact oligosaccharides from specific glycoproteins by endo–beta–N–acetylglucosaminidase H. *J. Biol. Chem.* **249**, 811–817.

Vriend, G. (1990). WHAT IF: a molecular modelling and drug design program. *J. Mol. Graph.* **8**, 52–56.

Wang., B.C. (1985). Resolution of phase ambiguity in macromolecular crystallography. *Meth. Enzymol.* **115**, 90–112.

Zhang, K.Y.J. & Main, P. (1988). Histogram matching as a density modification technique for phase refinement and extension of protein molecules. In *Improving protein phases.* Bailey, S., Dodson, E. & Phillips, S.E.V., eds., pp. 57–64., SERC Daresbury Laboratory, Warrington, UK.

Zeelen, J.Ph., Hiltunen, J.K., Ceska, T.A. & Wierenga, R.K. (1994). Crystallisation experiments with 2–enoyl–CoA hydratase, using an automated 'fast–screening' crystallisation protocol. *Acta Crystallogr.* **D50**, 443–447.

# Chapter 3

# Structure analysis of S1 nuclease

## 3.1 Quality of the model

The model of S1 nuclease has been refined to 1.7 Å resolution. The present model contains two NCS–related monomers per asymmetric unit linked by two zinc ions coordinated from both subunits. All side chain positions are unambiguously defined in the density map, although the position of the last three residues in molecule A and the last two residues in molecule B are undetermined. Each subunit binds three closely placed zinc ions in their active centre. Two additional zinc ions facilitate crystal contacts between symmetry related molecules. Each monomer has two N–linked carbohydrate side chains of which one (in molecule A) is completely missing from the electron density probably due to different molecular environment in the crystal structure. Almost 600 water molecules are built into the model, from which fifty have a B–factor less than ten. The waters with the lowest B values are in the coordination sphere of zinc ions and in the active site pocket.

A comparison of the $C_\alpha$ positions between the two NCS–related molecules (Figure 3.1) shows that the two peptide chains, with an overall r.m.s.$\Delta$ of 0.209 Å, are essentially identical despite the fact that no NCS restraints were used in the final stage of the refinement. Four residues, 82, 104, 181 and 228 deviate the most in their $C_\alpha$ positions. The main reason is the entirely different molecular environment or packing of residues 82, 181 and 228 in the respective NCS–related molecules, while residue 104 is in the worst determined part of the polypeptide chain. Molecule A and B are almost identical in terms of the B–factors of the main–chain atoms (Figure 3.2). A short loop from residue 99 to 105 has relatively high B–factor. This region has the poorest quality of the electron density

map. The residues of this loop also give the highest r.m.s.Δ when the $C_\alpha$ traces of S1 and P1 nucleases are superimposed on each other (Figure 3.4) and the only gap in the S1 nuclease sequence, shown by an alignment with P1 nuclease sequence (Appendix A), falls into this loop. The last C–terminal residues (residues 265, 266, 267 in molecule A and residues 266 and 267 in molecule B) do not show up in the electron density map at all. The increasing disorder towards the C–terminus is reflected by the increasing B–factors of the main–chain atoms in the modelled C–terminal residues (Figure 3.2).

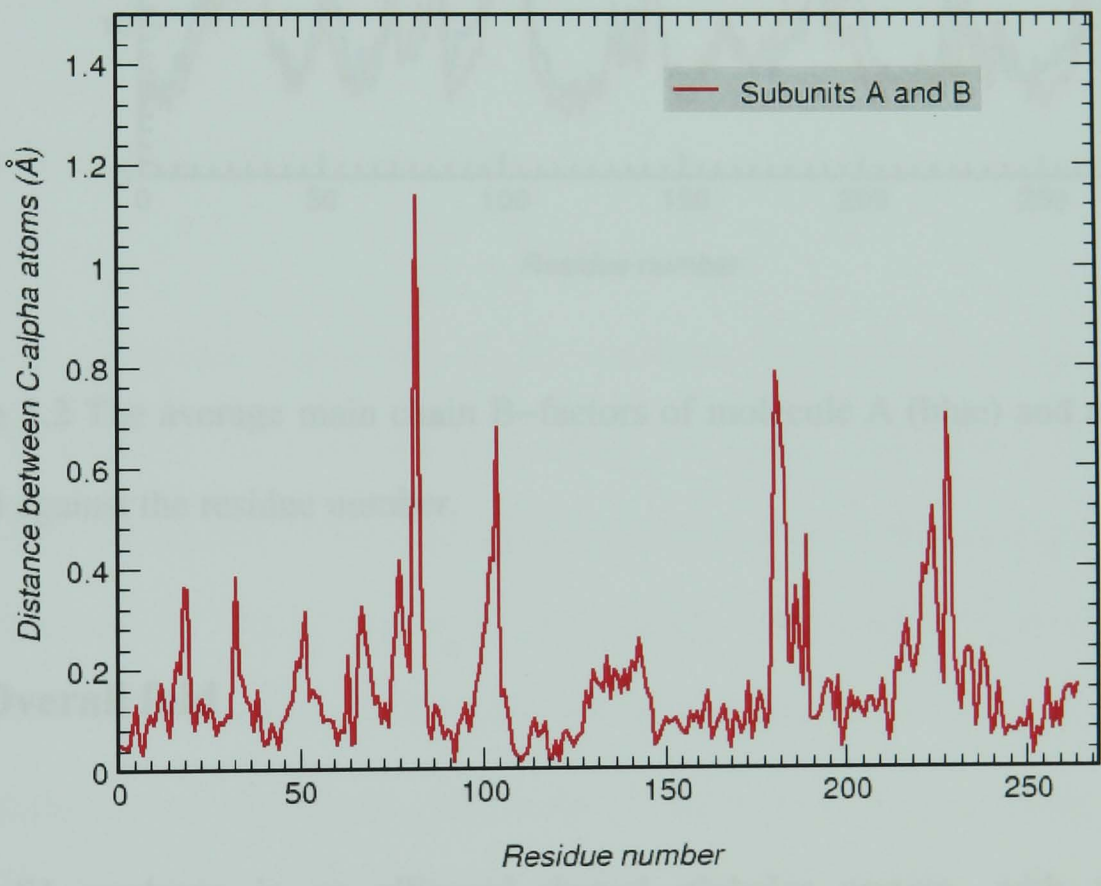



**Figure 3.1** The distances between equivalent $C_\alpha$ atoms of molecule A and B plotted against the residue number. The average r.m.s.Δ is 0.209 Å.

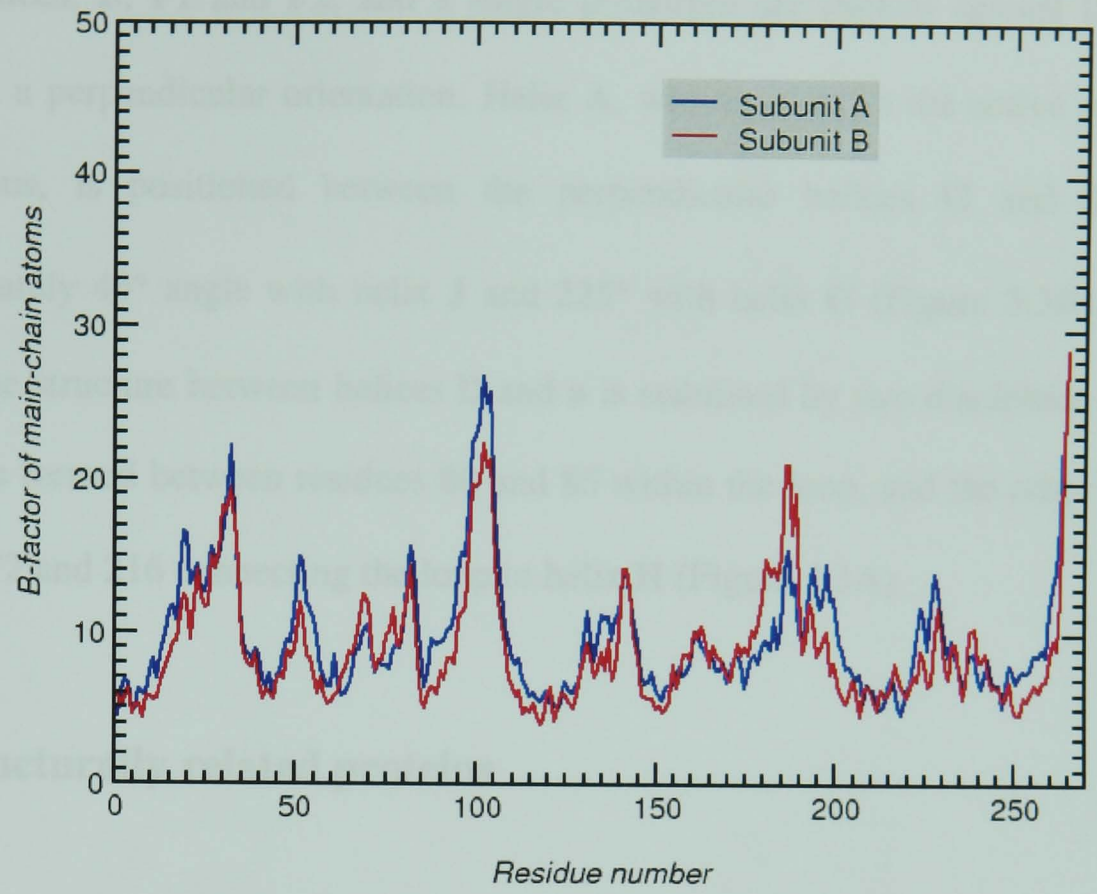



**Figure 3.2** The average main chain B–factors of molecule A (blue) and molecule B (red) plotted against the residue number.

## 3.2 Overall fold

S1 nuclease is an ellipsoid–shaped globular protein, with two protrusions perpendicular to the long axis of the molecule. Using a secondary structure assignment algorithm implemented in the program STRIDE (Frishman & Argos, 1995) ten α–helices, a $3_{10}$–helix and a β–hairpin can be distinguished in the structure. The helical content of the molecule is as high as 60%. Two pairs of antiparallel helices, **E–D** and **H–J** orient along the long axis of the molecule forming an approximately anti–parallel four–helix bundle (Figure 3.3B). One of the protrusions is formed by helix **G** and the loop connecting it to helix **H** on one side of the molecule, while the other protrusion is formed by helix **C** and

the $3_{10}$–helix **a** on the opposite side. Helix **I** connects the anti–parallel helix–pair **H–J**. Three helices, **B**, **F1** and **F2**, and a single β–hairpin are packed against the four–helix bundle in a perpendicular orientation. Helix **A**, which points to the active centre with its N–terminus, is positioned between the perpendicular helices **G** and **J** making an approximately 45° angle with helix **J** and 225° with helix **G** (Figure 3.3B). The longest loop in the structure between helices **D** and **a** is stabilised by two disulphide bridges. One of them is formed between residues 80 and 85 within the loop, and the other one between residues 72 and 216 connecting the loop to helix **H** (Figure 3.3A).

## 3.3 Structurally related proteins

The structural similarity of P1 nuclease from *Penicillinum citrinum* and phospholipase C (PLC) from *Bacillus cereus,* including their active site, was already known when the present thesis work began (Volbeda *et al.*, 1991). A high, 50% sequence identity between S1 and P1 nucleases also strongly suggested close structural similarity between S1 and P1 nucleases. Actually, a DALI search (Holm & Sander, 1993) in the FSSP database using the atomic coordinates of molecule B of S1 nuclease revealed three protein structures similar to the structure of S1 nuclease. Besides P1 nuclease and PLC, the structure of alpha–toxin from *Clostridium perfringens* (Naylor *et al.*, 1998) shows significant structural similarity to S1 nuclease. Alpha–toxin is a two domain protein, of which the larger N–terminal one is similar to S1 and P1 nucleases, and almost identical to PLC (Figure 3.4). Although PLC and alpha toxin are structurally similar to S1 nuclease, they share little similarity at the sequence level. Both PLC and alpha–toxin have only 16% sequence identity compared to S1 nuclease within 200 and 184 superimposed residues respectively.

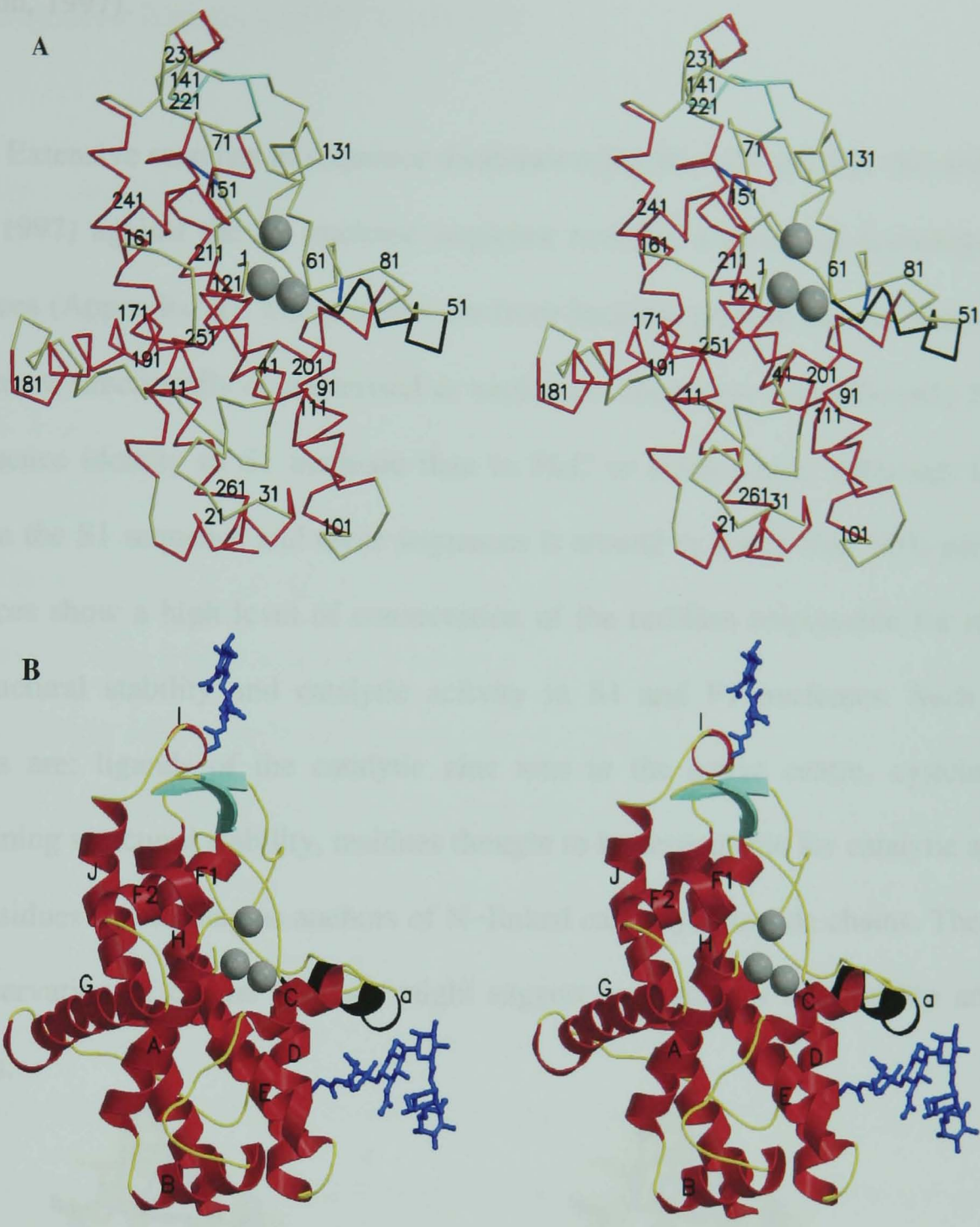



**Figure 3.3 A)** Stereo picture of the $C_\alpha$ trace of S1 nuclease with sequence numbering. α–helices are shown in red, the $3_{10}$–helix in black, the β–hairpin in aquamarine and the rest is shown in yellow. Helices are only assigned if they are formed by at least four residues. The active centre is marked by three zinc ions in grey; the two disulphide bridges are indicated by blue lines connecting $C_\alpha$ atoms. **B)** Secondary structure elements in S1 nuclease. The colouring scheme is the same as in A, except that the disulphide bonds are not drawn, but the two glycosylic side chains are shown as they are modelled in molecule B. The α–helices are marked with capital letters, the $3_{10}$–helix with letter "a". The figures were prepared with the programs MOLSCRIPT (Kraulis, 1991) and RASTER3D (Meritt

& Bacon, 1997).

Extensive searches in sequence databases using BLAST and PSI–BLAST (Altschul *et al.*, 1997) against the S1 nuclease sequence revealed a dozen of homologous protein sequences (Appendix A). The proteins are from bacteria, protozoans and plants, and most of them are functionally characterised as nucleases. They have a significantly higher level of sequence identity to S1 nuclease than to PLC or alpha–toxin. Although the identity between the S1 sequence and these sequences is around or lower than 30% percent, these sequences show a high level of conservation of the residues responsible for maintaining the structural stability and catalytic activity in S1 and P1 nucleases. Such important residues are: ligands of the catalytic zinc ions in the active centre, cysteine residues maintaining structural stability, residues thought to be responsible for catalytic activity and even residues functioning as anchors of N–linked carbohydrate side chains. The high level of conservation of crucial residues might suggest evolutionary relationship among these proteins.

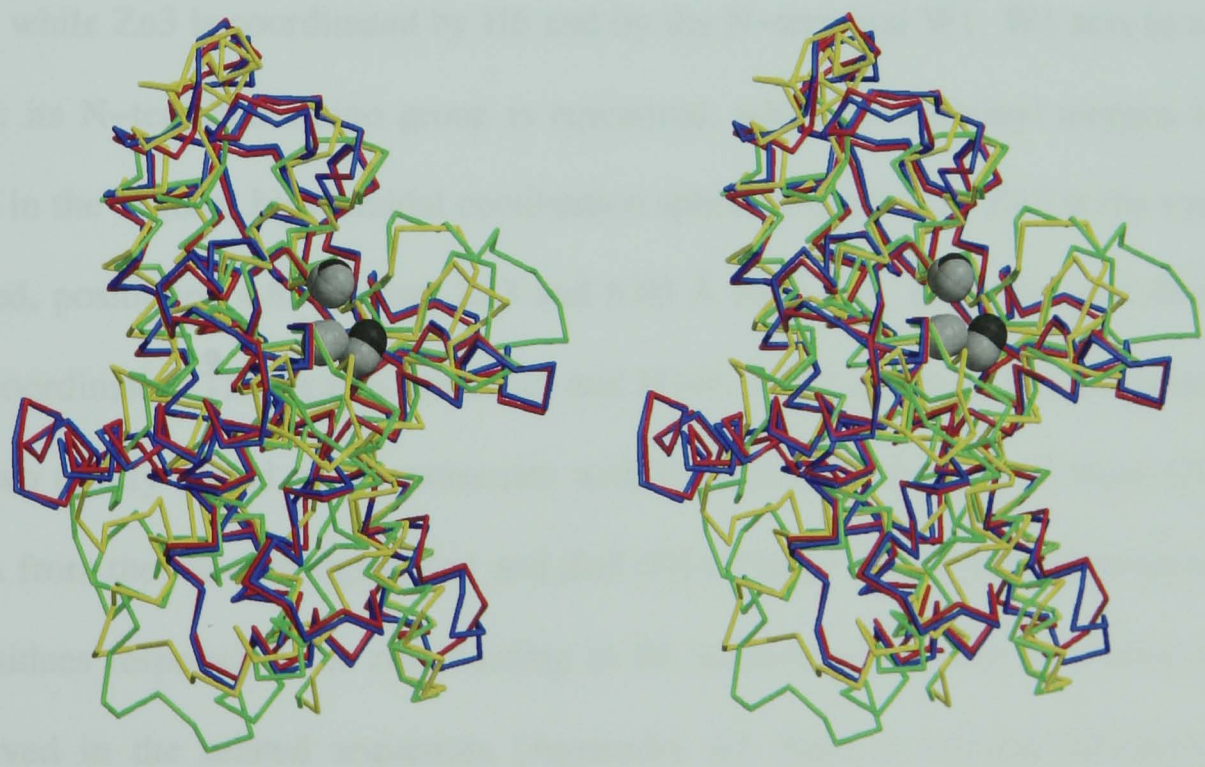


**Figure 3.4** Superposition of S1 (red) and P1 (blue) nuclease, PLC (gold) and alpha–toxin (green). The superposition was made by the DALI server using only related sequence segments.

## 3.4 Structural features of S1 nuclease

### 3.4.1 Zinc coordination

As discussed in chapter 1, S1 nuclease is a zinc dependent enzyme, and its enzymatic activity is completely abolished by removing the zinc ions from the molecules. Altogether twelve zinc ions per asymmetric unit have been identified and built into the solvent flattened electron density map calculated immediately after molecular replacement. Two classes of zinc ions bound in S1 nuclease can be distinguished: one is the trinuclear zinc cluster in the active site of the protein, the zinc ions of the other class facilitate crystal contacts. At the bottom of the active site cleft two zinc ions are 3.3 Å apart bridged by D119 as well as by a water (or rather hydroxide) molecule (O1). This water, has an unusually low temperature factor of 5 $Å^2$. Both Zn1 and Zn3 have five ligands in a distorted trigonal bipyramidal arrangement. The other ligands of Zn1 are D45, H60 and H115, while Zn3 is coordinated by H6 and by the N–terminal W1. W1 acts as a bidentate ligand: its N–terminal amino group is equatorial, while the carbonyl oxygen is an axial ligand in the trigonal bipyramidal coodination sphere (Figure 3.5). Zn2 is the most solvent exposed, positioned 4.85 Å from Zn3 and 6.03 Å from Zn1. Like Zn1 and Zn3 it is also pentacoordinated. The ligands are H125 and H148, D152 acting as a monodentate ligand, plus two tightly bound water molecules with B–factors ~6 $Å^2$, one of them (O2) is only 2.66 Å from the water bridging Zn1 and Zn3 (O1) (Figure 3.6). It is interesting to note that the residues responsible for zinc binding in S1 nuclease are totally or almost completely conserved in the related sequences (Appendix A). Such a striking similarity of these residues strongly suggests the presence of a trinuclear zinc cluster in the related proteins and predicts a very similar mechanism of action as well.

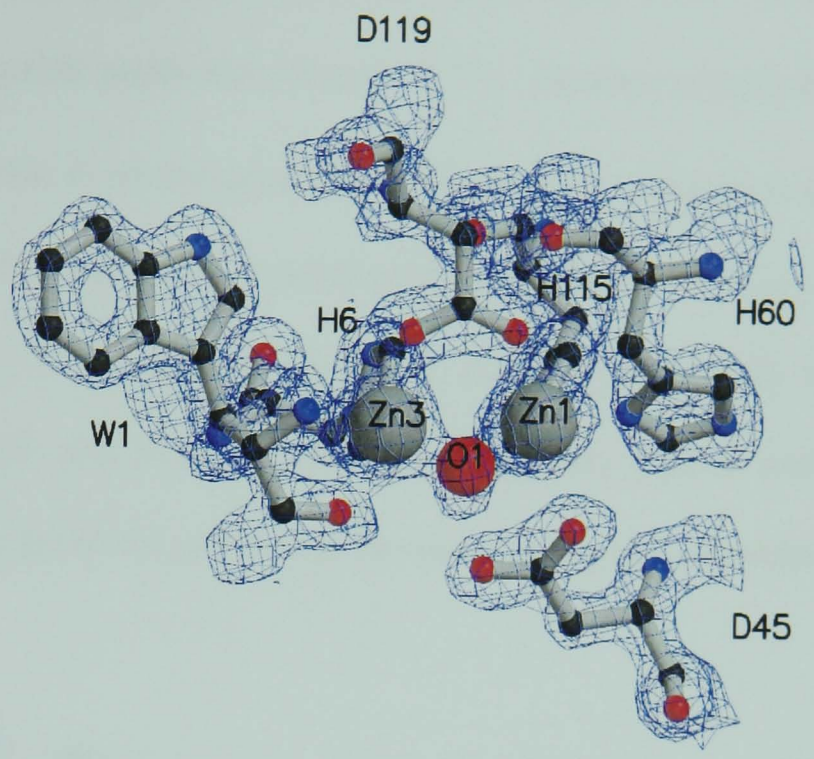



**Figure 3.5** The coordination sphere of the two closest placed zinc ions (3.3 Å), Zn1 and Zn3. These two ions are bridged by a water (or rather hydroxide) molecule O1. Note, that the plane of the imidazole rings of two coordinating histidine residues, H6 and H115 are just perpendicular to the plane of the paper and they are somewhat hidden by Zn1 and Zn3.

Zn4, Zn5 and Zn6 belong to another class of the zinc ions which are not involved in catalysis. Zn4 has a tetragonal bipyramidal coordination sphere. The ligands are D75 and D77 acting as monodentate equatorial ligands. The coordination sphere is completed by four water molecules, three of which are in hydrogen bonding contact with E7 from the same symmetry related molecule. Since the "external" ligands of Zn4 are only water molecules, therefore it is very probable that D75 and D77 bind zinc or similar bivalent metal ions also in solution. Zn5 is bound at the interface of the NCS–related molecules tetrahedrally coordinated by three residues, D221, D222 and K229 from one monomer,

whereas the fourth ligand, E235 is presented by the other NCS–related molecule. This kind of cross–coordination of Zn5 in both subunits creates a two–fold rotational symmetry between the two protein molecules (Figure 3.7A). Another remarkable feature of Zn5 is the participation of the ε–amino group of K239 in the coordination sphere (Figure 3.7B), which is unexpected at the pH of crystallisation (8.0), however it is well known that zinc has a good affinity to N–ligands (Valle & Auld, 1990a). Zn6 is also tetrahedrally coordinated like Zn5, and links together two symmetry related molecules of the same subunit. The ligands are H149 and D199, the latter acting as a monodentate ligand.

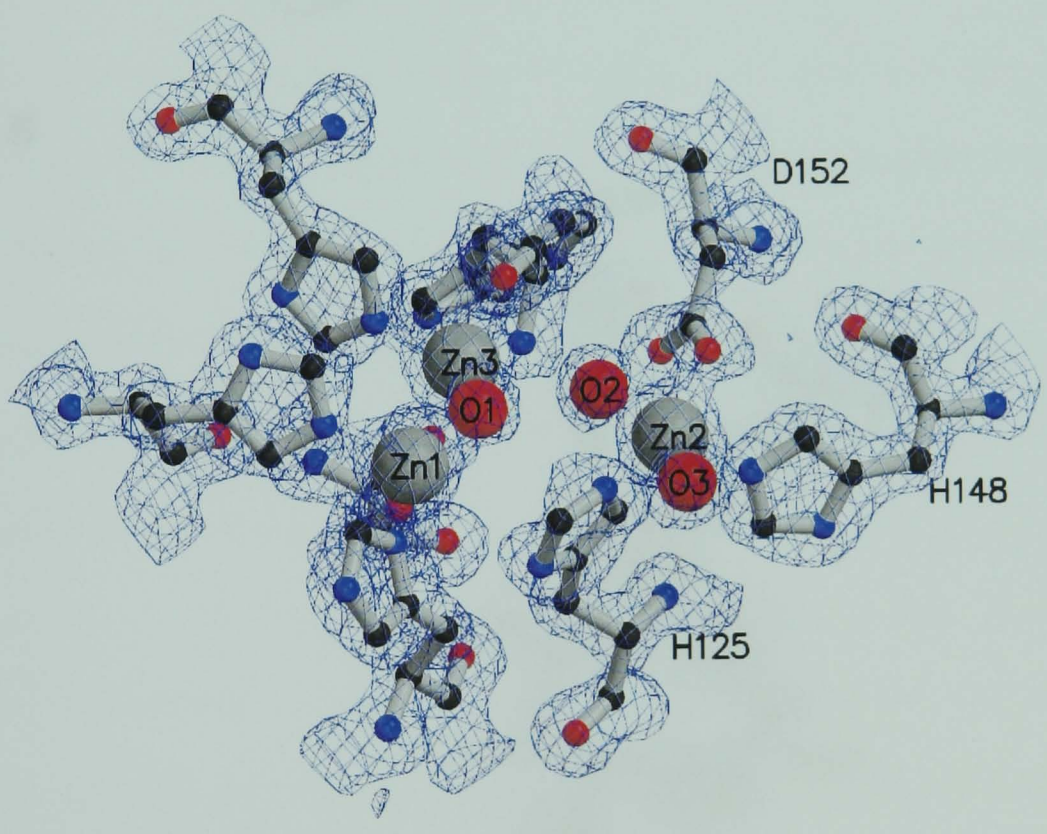
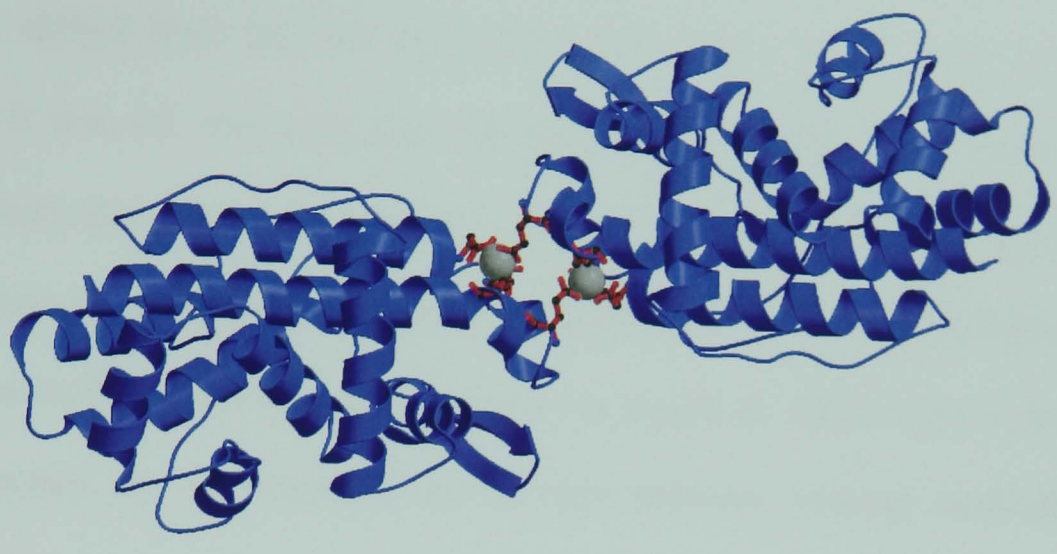



**Figure 3.6** The trinuclear zinc cluster. The coordination sphere of Zn2 shown together with Zn1 and Zn3.

**A**



**B**

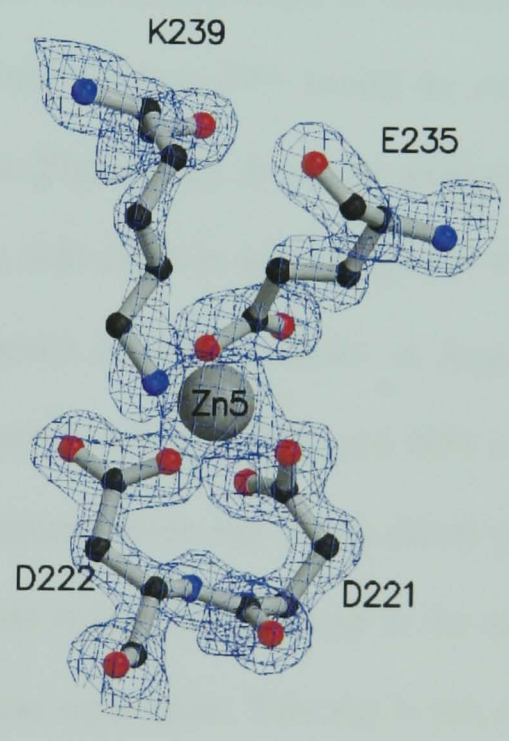



**Figure 3.7 A)** The coordination of Zn5 creates a non–crystallographic two–fold symmetry in the asymmetric unit. **B)** An unusual ligand, K239 in the coodination sphere of Zn5. The lysine residue must not be protonated to be able to interact with zinc, whereas it is supposed to be still protonated at pH 8.

## 3.4.2 Carbohydrate side chains

The apparent molecular weight of S1 nuclease (~32 kDa, see Figure 2.3) significantly differs from the MW calculated on the basis of the protein sequence (~29 kDa). Thus it was not surprising that extended density features have been found in the electron density map stemming from N92 and N228, which were attributed to N–linked sugar side chains in agreement with the biochemical data (Shishido & Habuka, 1986). The quality of the carbohydrate density as well as the length of chains built in the two NCS–related molecules are not identical due to their different packing environment in the crystal. In molecule A two *N*–acetyl–β–glucosamin moieties which are packed against several aromatic and hydrophobic side chains could be modelled into the density. Without the shielding by the carbohydrate residues F55 would be completely solvent exposed, which is entropically unfavoured (Figure 3.8). A similar arrangement has been found in P1 nuclease where the carbohydrate side chain is packed against W55 (Volbeda *et al.*, 1991). According to the multiple sequence alignment shown in Appendix A, there is a strong conservation of the residues corresponding to F55 and N92 in S1 nuclease. One might expect, therefore, N92 to be a glycosylation site in the whole protein family. The second glycosylation site at N228 in chain A does not show up in the electron density, the density is truncated at the ND2 atom of the asparagine. This site is not conserved in the S1 protein family. Molecule A and B have different packing environments and this is reflected in the appearance of the carbohydrate side chains in the density map. At residue N92 in chain B the first two *N*–acetyl–β–glucosamine moieties have the same environment as in chain A, but three more carbohydrate moieties can be observed as a continuation of the chain in molecule B. The third residue is a β–mannose, which is expected to be a branch point. Actually, there is a clear indication that the chain forks into two directions: from O3 and O6, but only one chain through an α–C1–O6–linkage to the next mannose can be clearly

followed. The fifth residue in the chain is also a mannose linked by an α–C1–O3 glycosyl bond to the previous mannose moiety (Figure 3.8). The identity of the mannose residues can be unambiguously determined due to their axial 2–hydroxyl group, which shows up nicely in the density. The presence of a mannose as the fifth member of the chain is surprising, since this position is occupied mostly by *N*–acetyl–β–glucosamine acting as another branch point. At N228 in molecule B, in contrast to molecule A, two *N*–acetyl–β–glucosamine residues linked by β–C1–O4–glycosyl bonds can be identified. As mentioned above here the sugar residues are packed against the protein, while in molecule A N228 is completely solvent exposed.

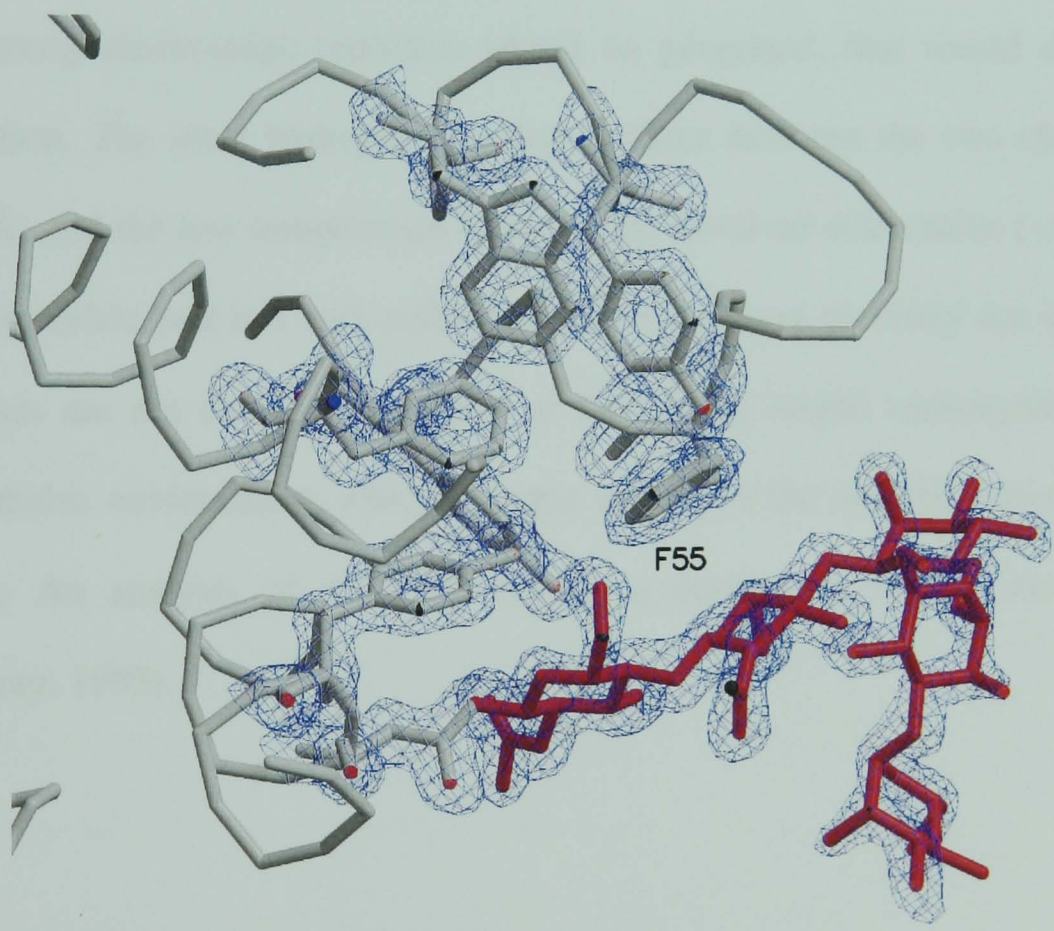



**Figure 3.8** The longest visible carbohydrate side chain in molecule B (red). The side chain is packed against a patch of hydrophobic aromatic residues in the same manner as in P1 nuclease. The phenyl ring in F55 is stacked with the second sugar residue in the carbohydrate side chain. In P1 the corresponding residue is W55 (Volbeda *et al.*, 1991).

### 3.4.3 Interacting carboxylates

A remarkable feature of the S1 nuclease model is a pair of buried interacting carboxylate side chains of D66 and E127 (Figure 3.9). Among the related sequences in Appendix A only P1 nuclease is known to have interacting buried carboxylate side chains at equivalent positions. In the other sequences only D66 is conserved, while in P1 there is an additional such pair, where D146 interacts with D151. The $pK_A$ of a solvent exposed carboxylate side chain is around 4.5, while the pH of crystallisation was 8.0. In case of solvent exposed carboxylate groups such a difference of $pK_A$ and the actual pH makes it improbable that the carboxylate groups are protonated. However, in the charged state a very strong electrostatic repulsion would be generated, that would destabilise such an interaction. The ideal hydrogen bonding distance between the two closest oxygen atom (2.46 Å) and the low temperature factor of the involved side chains (~6 $Å^2$) suggests that the carboxylate pair is a very stable formation and most probably not charged. S1 and P1 nucleases are not the only examples of interacting buried carboxylate side chains in a hydrophobic environment, which strongly influences the real pK values of acidic or basic groups. An analysis of the PDB reveals a number of similar examples (Flocco & Mowbray, 1995).

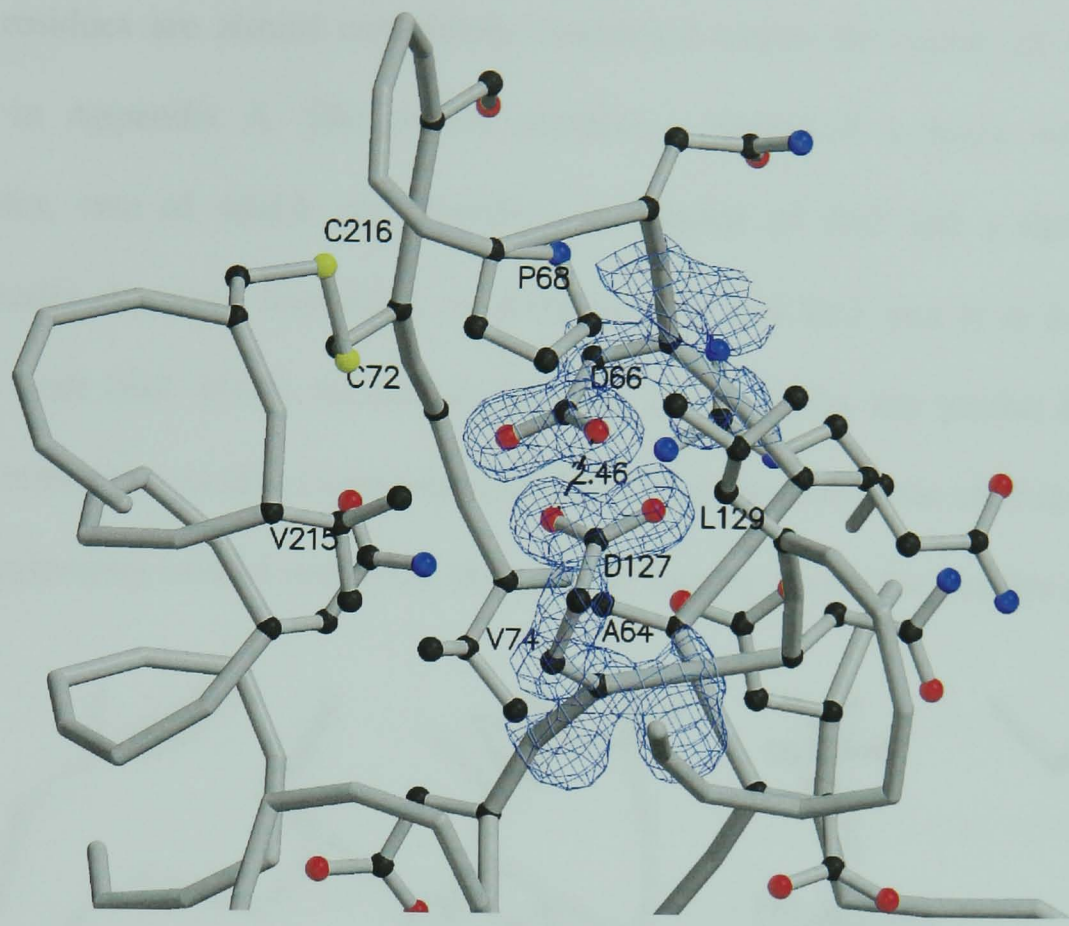



**Figure 3.9** Two interacting carboxylates in S1 nuclease surrounded by mostly hydrophobic residues.

## 3.5 The active site pocket in S1 nuclease

The active site of S1 nuclease can be easily identified by locating the trinuclear zinc cluster in the molecule. Most of the residues forming the active site pocket function as ligands of the zinc atoms. The crucial role of the zinc coordinating residues was shown earlier. Chemical modification of carboxylate and imidazole groups led to the loss of zinc atoms, and as a consequence to the complete loss of catalytic activity (Gite & Shankar, 1992a; Gite & Shankar, 1992b) The residues acting as zinc ligands are W1 and H6 from helix **A**, D45 from helix **C**, H60 from the $3_{10}$ helix **a**, H115 and H119 from helix **E**, H125 from the loop connecting helices **E** and **F**, and finally H148 and D152 from helix **F**. Further residues of the active site pocket not involved in metal coordination are K48 and Y49 from helix **C**, and the solvent exposed F61, the first residue in the loop connecting

helices **a** and **D,** which has a relatively weak density in the $2F_o-F_c$ map (Figure 3.10). These residues are almost completely conserved within the family of related sequences shown in Appendix A. The pocket contains a cluster of a dozen well defined water molecules, two of which are identified as ligands of Zn2 and a third one, which is catalytically the most important, is bridging Zn1 and Zn3 and is in hydrogen bonding contact with D45. H149, located at the edge of the active site pocket links a symmetry related molecule via the coordination of Zn6. Thus the active site cleft is partially blocked by the symmetry related molecule, though still leaving access for smaller substrates.

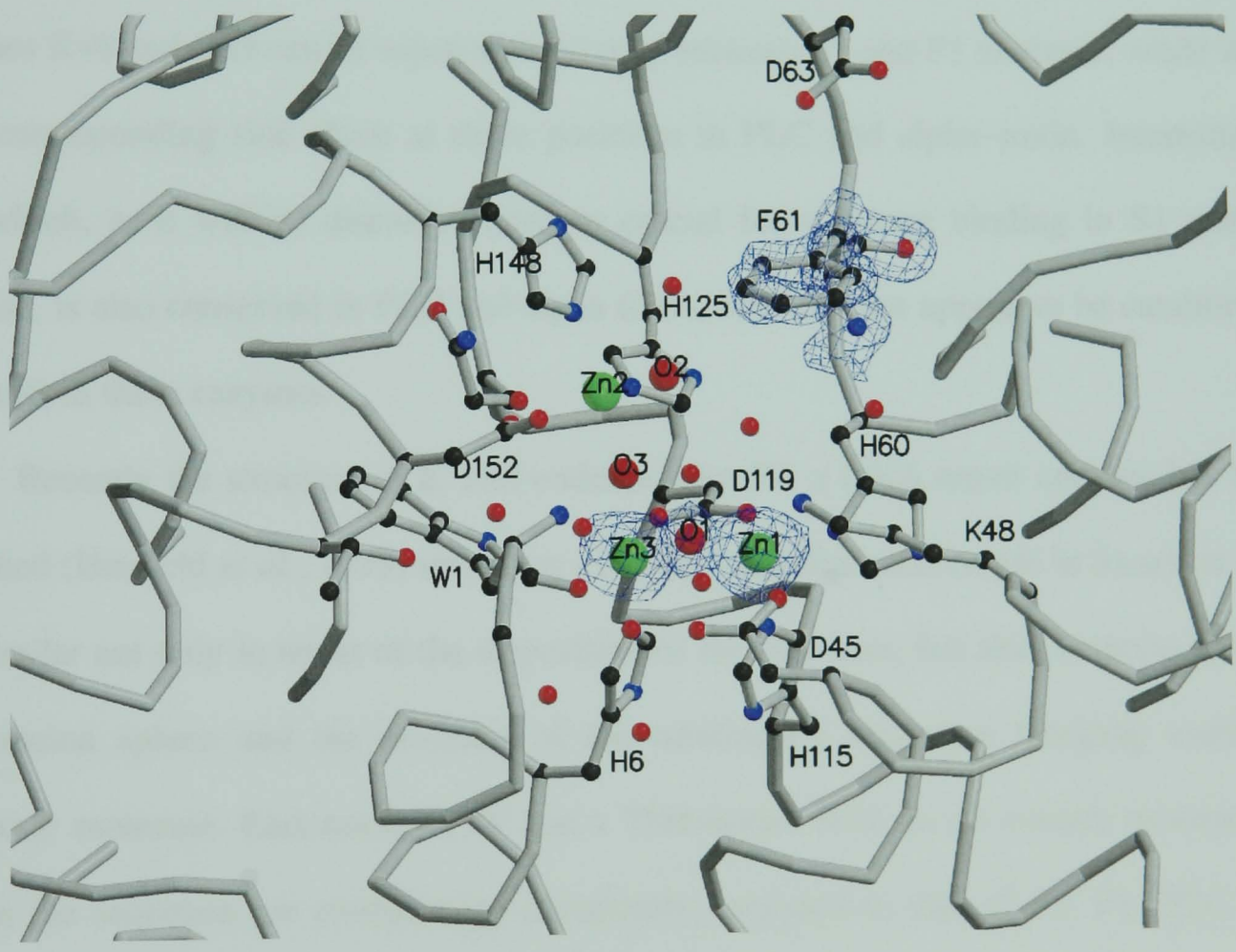



**Figure 3.10** The active site in S1 nuclease. The trinuclear zinc cluster and the catalitically important residues (including) waters are labelled.

## 3.6 Comparison of the active sites of enzymes with a trinuclear zinc cluster

The structures discussed in chapter 3.3 not only share a similar overall fold, but they also possess the same trinuclear zinc cluster. When the structures are superimposed to obtain the best fit of the $C_\alpha$ positions, then the zinc atoms and their ligands are also closely overlapping (Figure 3.11). The zinc ligands are essentially identical, except D152 in S1 nuclease, which is replaced by E146 and E152 in PLC and alpha–toxin respectively. Residues K48 and D63 can be superimposed only between S1 and P1 nuclease, while there is no corresponding side chain at those positions in PLC and alpha–toxin. Interestingly F61, which, as it will be discussed later, is crucial for substrate binding in S1 and P1 nuclease, is also conserved in PLC and alpha toxin, but does not appear to be catalitically important in these enzymes.

Recently the structure of *E. coli* endonuclease IV, a DNA repair enzyme has been published (Hosfield *et al.*, 1999) revealing a similar trinuclear zinc centre in its active site. It is similar not only in terms of the disposition of the zinc ions, but also in terms of their coordination sphere and the presence of the catalitically important bridging water or hydroxide molecule. Endonuclease IV has a TIM–barrel fold, so its overall structure as well as the sequence are evolutionary completely unrelated to that of S1, P1, PLC and alpha–toxin, while on the other hand it is expected to cleave the P–O bond via the same catalytic mechanism. The similarity of the active site of these enzymes may represent an interesting new example of convergent evolution (Russell, 1997).

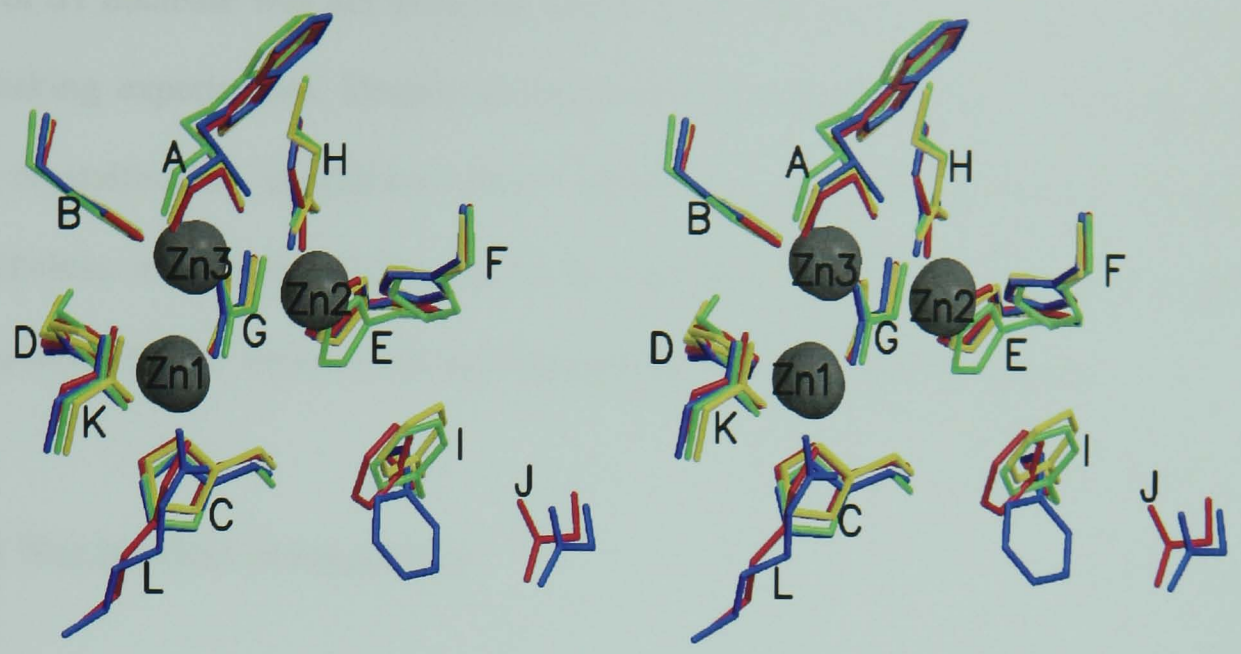



**Figure 3.11** Superposition of active centre residues of PLC (gold), alpha–toxin (green), P1 nuclease (blue) and S1 nuclease (red). Residues labeled with **L** and **J** are only shown for S1 and P1 nucleases. The side chain of F61 in P1 nuclease (label **I**) is obviously deviating from the others, because the substrate bound high resolution structure was used for superposition. The residues contributing to zinc coordination are identical except the residue labelled **H**, where aspartic acid is replaced by glutamic acid in PLC and alpha–toxin.

## 3.7 Proposed mechanism of action in S1 nuclease: the three–metal ion mechanism

One of the goals of the present work was to elucidate the enzymatic mechanism of S1 nuclease and to explain the differences in substrate preferences between S1 and P1 nucleases. In order to study the interactions in enzyme–substrate complexes one has to use either a mutant enzyme capable of binding but unable to cleave or uncleavable substrate analogues. Although in both cases a non–productive complex would be obtained, it might provide enough information to draw conclusions about the catalytic mechanism. Since the

clone of S1 nuclease was not available substrate analogues were used in co–crystallisation and soaking experiments. Despite several trials no complex crystals have been obtained using crystallisation conditions where diffraction quality S1 nuclease crystals form. Nevertheless, a catalytic mechanism of S1 nuclease can be proposed based mainly on a comparison with the structurally and functionally very similar P1 nuclease.

### 3.7.1 Nucleotide recognition

P1 nuclease possesses an identical fold, and an almost identical active site compared with S1 nuclease. As a result, the conclusions drawn from the structural analysis of a P1–substrate analogue complex can be safely applied to S1 nuclease too. Soaking P1 nuclease with the uncleavable *R* diastereomer of Ap(S)A (Potter *et al.*, 1983), a phosphoromonothioate was the first attempt to obtain enzyme–substrate complexes (Volbeda *et al.*, 1991) followed by co–crystallisation with phosphorodithioates (Romier *et al.*, 1998). Although soaking with *R*–Ap(S)A did not result in a refined model due to the low resolution of the diffraction data (~4 Å), it allowed the identification of two nucleotide binding sites, into which a 5'–AMP molecule could be modelled. The crystal structure of a P1–ATTT phosphorodithioate complex (Romier *et al.*, 1998) also revealed a similar non–productive binding mode of the oligonucleotide analogue, but showed much more details at 1.8 Å resolution. In this case the oligonucleotide links together two molecules in the crystal structure such that its 5' end binds to the second binding site of one P1 molecule, while the 3' end binds to the primary binding site of another P1 monomer. The two middle residues in the tetranucleotide don't make any contacts with the protein. The manner of nucleotide recognition is basically the same as seen in the low resolution P1–*R*–Ap(S)A complex. The first nucleotide binding site involves F61, which in turn is very close to the trinuclear zinc cluster (Figure 3.10 & 3.12). The base of the nucleotide is stacking onto the

benzene ring of F61, and forms hydrogen bonding contacts with D63 (Figure 3.10 & 3.12). It is important to note , that in order to establish optimal hydrogen bonding interactions with the nucleobase, D63 has to be protonated in which case it can act as hydrogen bond donor as well as acceptor (see Romier *et al.*, 1998). This fact might explain the failure to co–crystallise oligonuclotide substrates with S1 nuclease at relatively high pH (8.0), when D63, being solvent exposed, is most probably deprotonated. The importance of the stacking interaction between F61 and nucleotide base has been experimentally demonstrated. P1 and S1 nucleases are unable to cleave phosphodiester bonds with an abasic 5' nucleotide, while a 3' abasic nucleotide has no effect on the cleavage efficiency (Weinfield *et al.*, 1989). In addition, oligonucleotides with ring–saturated base analogues in the 5' position were found to be weak substrates or not to be substrates of S1 and P1 nucleases (Weinfeld *et al.*, 1993). The second nucleotide binding site, which is missing in S1 nuclease, is 20 Å away from the active site and is formed by two closely spaced tyrosine residues. In P1 the base of one nucleotide is stacked between the two tyrosyl side chains. The role of this second nucleotide binding site in P1 which is far from the active centre and is absent from S1 nuclease, is unclear.

### 3.7.2 Catalytic mechanism

The structure of the P1/*R*–Ap(S)A complex provided the first clues on the possible reaction mechanism of P1 nuclease. Three different reaction mechanisms were proposed, in agreement with the evidence that the cleavage of the phosphodiester bond proceeds with inversion of configuration at the phosphorous (Potter *et al.*, 1983) through a pentacovalent transition state. A zinc activated water molecule has been proposed as the attacking nucleophile and R48 of P1 (K48 for S1), a positively charged residue, was thought to stabilise the transition state. However, it was not clear which of the zinc activated water

molecules acts as a nucleophile. In one mechanism the water molecule bridging Zn1 and Zn3 acts as the nucleophile, while in another mechanism a water molecule coordinating the more exposed Zn2 is proposed (Figure 3.5 & 3.6). In a third mechanism a phosphate oxygen is bound between Zn1 an Zn3 replacing the bridging water molecule. The latter mechanism is analogous to that proposed for the 5'–3' exonuclease activity of *E. coli* DNA polymerase I, known as the "two–metal ion mechanism" (Beese & Steitz, 1991).

The high resolution structure of P1–$Ap(S)_2Tp(S)_2Tp(S)_2T$ complex provided more information on the possible catalytic mechanism, even though it is not a productive complex. In this complex the 3'–terminal thymine of ATTT is stacked against F61 and forms hydrogen bonding contacts with D63, while the O3' hydroxyl of the deoxyribose is bound to Zn2 replacing one of its strongly bound waters. The complex can be considered an enzyme–product complex just after the cleavage of the 3'–terminal part of the substrate. If the chain is extended in the 3' direction, then the scissile phosphate is positioned between the three zinc ions close to Zn2 replacing one of the strongly bound water molecules by one of its non–bridging oxygens, while the other non–bridging oxygen interacts with R48 of P1 (K48 for S1 nuclease). In such an arrangement the bridging water between Zn1 and Zn3, which is rather a hydroxide ion due to the electrophile nature of the zinc ions, is in–line with the O3'–P bond. D45, which is a ligand of Zn1 and is conserved within the related sequences (Appendix A), helps to properly orient the attacking hydroxide. The negatively charged pentacovalent transition state is stabilised by R48 of P1 (K48 for S1 nuclease), while the attacking hydroxide and the leaving O3' of the deoxyribose occupy apical positions. The leaving O3' is stabilised by coordination to Zn2, which as a Lewis–acid increases the electrophilicity of the phosphorous (Figure 3.12).

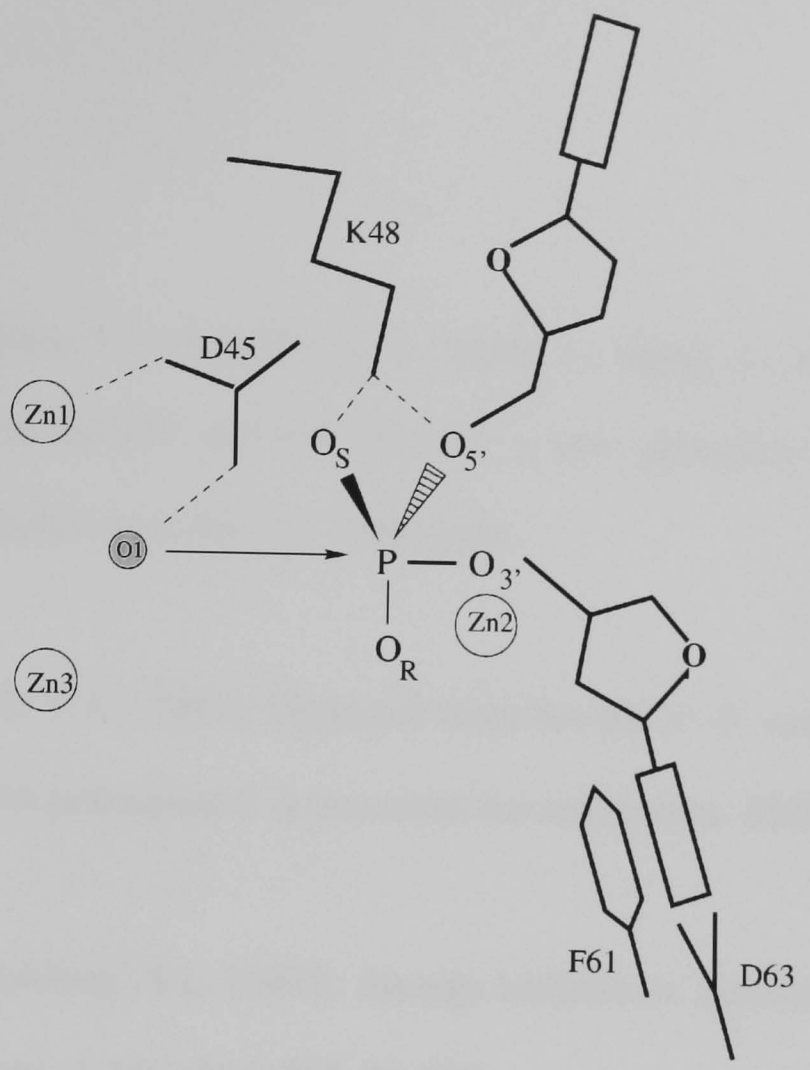



**Figure 3.12** The proposed mechanism of action for S1 nuclease and, in general, for hydrolases with trinuclear zinc cluster in the active site.

The proposed catalytic mechanism described above and schematically shown in Figure 3.12, involves all three zinc ions in the active site, making all of them essential for catalytic activity. Therefore the term 'three–metal ion' mechanism was proposed (Romier *et al.*, 1998). A similar three–metal ion mechanism was proposed for phospholipase C from *Bacillus cereus* based on computer simulations (Sundell *et al.*, 1994) and recently for endonuclease IV from *E. coli*. As described above, PLC has an almost identical active site geometry to S1 and P1, and possesses a similar all–helix fold (Hough *et al.*, 1989), whereas endonuclease IV only shares the trinuclear zinc cluster with a similar, but clearly identical coordination sphere. However, the conservation of the analogous residue D45 in S1 and P1 nucleases and the presence of the bridging hydroxide between Zn1 and Zn3

strongly suggests a similar three–metal ion mechanism (Hosfield *et al.*, 1999).

## 3.8 References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped BLAST and PSI–BLAST: a new generation of protein database search programs. *Nucleic Acid Res.* **25**, 3389–3402.

Beese, L.J. & Steitz, T.A. (1991). Structural basis for the 3'–5' exonuclease activity of *Escherichia coli* DNA polimerase I: A two metal ion mechanism. *EMBO J.* **10**, 25–33.

Flocco, M.M, & Mowbray, S.L. (1995). Strange bedfellows: interactions between acidic side–chains in proteins. *J. Mol. Biol.* **254**, 96–105.

Frishman, D. & Argos, P. (1995). Knowledge–based secondary structure assignment. *PROTEINS: Structure, Function and Genetics* **23**, 566–579.

Gite, S. & Shankar, V. (1992a). Active–site characterization of S1 nuclease I. *Biochem. J.* **285**, 489–494.

Gite, S. & Shankar, V. (1992b). Active–site characterization of S1 nuclease II. *Biochem. J.* **288**, 571–575.

Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138.

Hosfield, D.J., Guan, Y., Haas, B.J., Cunningham, R.P. & Tainer, J.A. (1999). Structure of the DNA repair enzyme endonuclease IV and its DNA complex: double–nucleotide flipping at abasic sites and three–metal–ion catalysis. *Cell* **98**, 397–408.

Hough, E. et al. & Derewenda, Z. (1989). High–resolution (1.5 Å) crystal structure of phospholipase C from *Bacillus cereus*. *Nature* **338**, 357–360.

Kraulis, P.J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950.

Merritt, E.A. & Bacon, D.J. (1997). Raster 3D photorealistic molecular graphics. *Methods in Enzymology* **277**, 505–524.

Naylor, C.E., Eaton, J.T., Howells, A., Justin, N. & Moss, D.S. (1998). Structure of the key toxin in gas gangrene *Nat. Struct. Biol.* **5**, 738

Potter, B.V.L., Connolly, B.A. & Eckstein, F. (1983a). Synthesis and configurational analysis of a dinucleoside phosphate isotopically chiral at phosphorus. Stereochemical course of Penicillinum citrinum nuclease P1 reaction. *Biochemistry* **22**, 1369–1377.

Potter, B.V.L., Romaniuk, P.J. & Eckstein, F. (1983b). Stereochemical course of DNA hydrolysis by nuclease S1. *J. Biol. Chem.* **258**, 1758–1760.

Romier, C., Dominguez, R., Lahm, A., Dahl, O. & Suck, D. (1998). Recognition of single–stranded DNA by nuclease P1: High resolution crystal structures of complexes with substrate analogs. *PROTEINS* **32**, 414–424.

Russell, R.B. (1997). Detection of protein three–dimensional side–chain patterns: new examples of convergent evolution. *J. Mol. Biol.* **279**, 1211–1227.

Shishido, K. & Habuka, N. (1986). Purification of S1 nuclease to homogeneity and its chemical, physical and catalytic properties. *Biochim Biophys. Acta* **884**, 215–218.

Sundell, S., Handen, S. & Hough, E. (1994). A proposal for the catalytic mechanism in phospholipase C based on interaction energy and distance geometry calculations. *Protein Eng.* **7**, 571–577.

Vallee, B.L. & Auld, D.S. (1990). Active–site zinc ligands and activated $H_2O$ of zinc enzymes. *Proc. Natl. Acad. Sci. USA* **87**, 220–224.

Volbeda, A., Lahm, A., Sakiyama, F. & Suck, D. (1991). Crystal structure of Penicillinum citrinum P1 nuclease at 2.8 Å resolution. *EMBO J.* **10**, 1607–1618.

Weinfeld, M., Luzzi, M. & Paterson, M.C. (1989). Selective hydrolysis by exo– and endonucleases of phosphodiester bonds adjacent to an apurinic site. *Nucleic Acids Res.* **17**, 3735–3745.

Weinfeld, M., Soderlind, K.–J.M. & Buchko, G.W. (1993). Influence of nucleic acid base aromaticity on substrate reactivity with enzymes acting on single–stranded DNA. *Nucleic Acids Res.* **21**, 621–626.

# Part B: The crystal structure of an Sm–related protein from *Archaeoglobus fulgidus*

## Chapter 4

## Introduction

An Sm related (or Sm–like) protein with presently unknown function has been cloned from the archaeon *Archaeoglobus fulgidus*, and its structure has been determined by X–ray crystallography. This structure and its implications will be discussed throughout the next three chapters. Sm and Sm–like proteins together with small nuclear RNA are the core components of several small nuclear ribonucleoprotein particles (snRNPs), which play essential roles in many aspects of gene expression. SnRNPs are involved in various cellular processes including pre–mRNA splicing (e.g. U1, U2, U4–U6 snRNPs), histone mRNA 3' end processing (U7 snRNP), rRNA processing (e.g. U3, U8, U13–72 snRNPs and RNase MRP), telomere replication (telomerase) and tRNA maturation (RNase P) (reviewed by Mattaj *et al.*, 1993). SnRNPs involved in pre–mRNA splicing (Lührman *et al.*, 1990), histone maturation (Smith *et al.*, 1991) and, as recently published, in telomere replication (Seto, *et al.*, 1999) contain the core Sm protein complex. From those the spliceosomal snRNPS are by far the best characterised. The recently determined X–ray structures of two Sm proteins from human show close structural homology with the Sm–related protein from *Archaeoglobus fulgidus*. In the following a short introduction will be given to the eukaryotic spliceosomal snRNPs and the Sm proteins constituting them.

## 4.1 Nuclear pre–mRNA splicing and spliceosome assembly

In eukaryotes most of the transcribed genes result in pre–mRNA which contain non–coding intervening sequences (introns), which have to be excised prior to translation into protein. The splicing reaction can be divided into two successive trans–esterification steps. The first step is a nucleophile attack by the 2'–hydroxyl group of a conserved adenosine within the intron region, known as *branch point*. In this reaction the 2'–hydroxyl esterifies the 5' splice site of the intron resulting in a circular lariat intron intermediate and in a free 5' exon. In the second step the 5' exon trans–esterifies the phosphodiester bond at the 3' splice site of the circular lariat intron intermediate resulting in the ligation of the two exons and the release of the circular intron. The later is first debranched by specific enzymes, then cleaved by RNases (Moore, Query & Sharp, 1993; Burge, Tuschl & Sharp, 1999; Baserga & Steitz, 1993; Staley & Guthrie, 1998).

The splicing reaction takes place in the cell nucleus, where the pre–mRNA first associates in an ordered manner with several proteins and complexes of proteins with small nuclear RNA (snRNA) forming an approximately 4.8 MDa catalytic unit called the *spliceosome* (Moore, Query & Sharp, 1993; Burge, Tuschl & Sharp, 1999; Müller *et al.*, 1998). The major components of the spliceosomes are snRNA–protein complexes termed small nuclear ribonucleoprotein particles (snRNP). Four such complexes: U1, U2, U4/U6 and U5 have been identified and named after their snRNA component. In the process of spliceosome assembly (Figure 4.1) U1 snRNP binds first to the conserved 5' splice site of the intron, followed by U2, binding to the branch point. Finally, the pre–assembled tri–snRNP, U4/U6·U5 joins the complex. In the spliceosome the original base pairing between U4 and U6 snRNA is interrupted, and a new network of interaction between U6 and U2 and between U6 and the 5' splice site forms (Moore, Query & Sharp, 1993; Burge, Tuschl & Sharp, 1999; Baserga & Steitz, 1993; Staley & Guthrie, 1998; Madhani & Guthrie,

1994). The U5 snRNP plays a crucial role in the second trans–esterification reaction, in which one of its conserved loops binds to both exons at their splice sites, keeping them spatially close (Newman & Norman, 1992; Sontheimer & Steitz, 1993; O'Keefe, Norman & Newman, 1996).

According to present knowledge, about 80–100 protein factors are involved in metazoan splicing, which can be divided into the group of snRNP associated proteins and the non–snRNP splicing factors. The non–snRNP splicing factors have been classified according to their function or sequence similarity to known proteins. They include various enzymes: ATPases, helicases, protein kinases, GTPases, peptidyl–prolyl cis/trans isomerases and others (Burge, Tuschl & Sharp, 1999; Will & Lührmann, 1997; Krämer, 1996; Beggs, 1995; Lührmann, Kastner & Bach, 1990; Nagai & Mattaj, 1994).

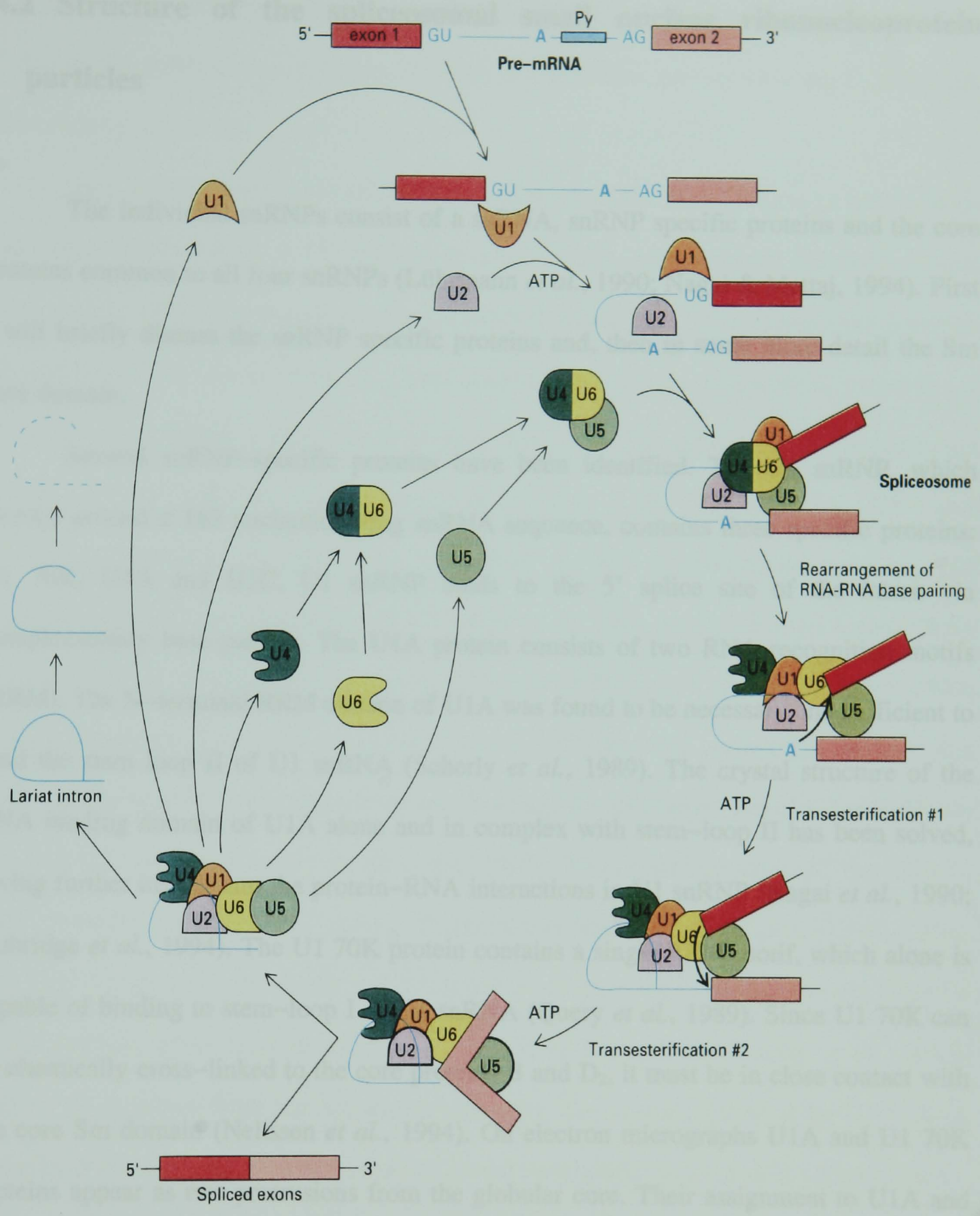



**Figure 4.1** The schematic drawing of the eukaryotic splicing cycle. The snRNPs bind on the branch point followed by a reorganisation step. After two successive transesterification reactions the spliced exons are released and the spliceosome disassembles. After debranching the lariat intron is degraded by various RNases.

## 4.2 Structure of the spliceosomal small nuclear ribonucleoprotein particles

The individual snRNPs consist of a snRNA, snRNP specific proteins and the core proteins common to all four snRNPs (Lührmann *et al.*, 1990; Nagai & Mattaj, 1994). First I will briefly discuss the snRNP specific proteins and, then in some more detail the Sm core domain.

Several snRNP–specific proteins have been identified. The U1 snRNP, which formed around a 163 nucleotide long snRNA sequence, contains three specific proteins: U1 70K, U1A and U1C. U1 snRNP binds to the 5' splice site of the intron via complementary base pairing. The U1A protein consists of two RNA recognition motifs (RRM). The N–terminal RRM domain of U1A was found to be necessary and sufficient to bind the stem loop II of U1 snRNA (Scherly *et al.*, 1989). The crystal structure of the RNA binding domain of U1A alone and in complex with stem–loop II has been solved, giving further insight into the protein–RNA interactions in U1 snRNP (Nagai *et al.*, 1990; Oubridge *et al.*, 1994). The U1 70K protein contains a single RRM motif, which alone is capable of binding to stem–loop I of U1 snRNA (Query *et al.*, 1989). Since U1 70K can be chemically cross–linked to the core proteins B and $D_2$, it must be in close contact with the core Sm domain (Nelissen *et al.*, 1994). On electron micrographs U1A and U1 70K proteins appear as two protrusions from the globular core. Their assignment to U1A and U1 70K respectively, was achieved using specific antibodies to the individual proteins (Kastner *et al.* 1992). The U1C protein does not bind directly to the U1 snRNA. It binds to U1 snRNP only in the presence of both the U1 70K protein and the core domain, suggesting that U1C binds directly to these two proteins. The latter assumption is supported by the observation of chemical cross–linking between U1C and core protein B (Nelissen *et al.*, 1991). The U1C protein is necessary for efficient complex formation

between the U1 snRNP and the 5' splice site, it is thought to alter the conformation of the 5' end of U1 snRNA, thus enabling efficient efficient base pairing with the 5' splice site (Heinrichs *et al.*, 1990).

As a second step the U2 snRNP binds to pre–mRNA in the course of spliceosome assembly. It is based on a 187 nucleotide long snRNA, which forms four stem–loops. Stem loop I facilitates interactions with U6 snRNA, when the original interactions between U6 and U4 snRNA are disrupted in a later stage of spliceosome assembly. A conserved, 6 nucleotide long stretch downstream from stem–loop I is responsible for base pairing to the branch point of the intron. In addition to the Sm core domain two proteins, U2B'' and U2A', were found as constituents of U2 snRNP at high salt conditions, whereas nine additional proteins are associated with U2 snRNP if the ionic strength is low. Among those nine are the heteromeric splicing factors SF3a and SF3b (Burge, Tuschl & Sharp, 1999; Will & Lührmann, 1997; Krämer, 1996). Studies with the yeast homologue of U2A' and U2B'' demonstrated, that both proteins are necessary for the integration of U2 snRNP into the pre–spliceosome (Caspary & Seraphin, 1998). The crystal structure of the U2B''–U2A' complex bound to an U2 snRNA fragment has been determined (Price *et al.*, 1998). The U2B'' protein is very similar to U1A at the sequence level, and binds to stem–loop IV of U2 snRNA, while U1A binds to stem–loop II of U1 snRNA (Scherly *et al.*, 1990). Indeed, as expected from the sequential similarity, the crystal structures of U1A–RNA complex and U2B''–U2A'–RNA complex reveal very similar protein–RNA interactions.

The structure of the U4/U6 snRNP is based on the extensively base paired snRNAs U4 and U6. The core domain, which has a globular shape on electron micrographs (Kastner *et al.*, 1990ab; Kastner *et al.*, 1991; Kastner 1998), is thought to contain the core Sm domain bound to U4 snRNA. U6 snRNA does not have an Sm binding site (Liautard *et al.*, 1982), however, Sm–like proteins with significant sequence similarity to the canonical Sm proteins have been found in yeast and man associated with U6 snRNA (Cooper *et al.*, 1995; Séraphin, 1995; Salgado–Garrido *et al.*, 1999; Achsel *et al.*, 1999).

As it was already mentioned U5 snRNP is crucial for the last trans–esterification step of splicing (Newman & Norman, 1992; Sontheimer & Steitz, 1993). U5 snRNP is fairly complex, it contains nine specific proteins in addition to the Sm core domain. Prior to joining the pre–spliceosome snRNPs U4/U6 and U5 associate to form a tri–snRNP complex U4/U6·U5 (Will & Lührmann, 1997).

In the context of this part of the thesis, the Sm core proteins of the U snRNPs have the most relevance. These proteins are common to all spliceosomal snRNPs except U6 (Burge, Tuschl & Sharp, 1999; Will & Lührmann, 1997). They have been found as the target of autoantibodies from patients suffering of the autoimmune disease systemic lupus erythmatosus (Lerner & Steitz, 1979). Seven canonical Sm proteins have been identified forming both the human (reviewed by Lührmann *et al.*, 1990) and the yeast Sm core domain (Rydmond, 1993; Roy *et al.*, 1995; Séraphin, 1995; Bordonne & Tarassov, 1996; Fromont *et al.*, 1997; Salgado–Garido *et al.*, 1999). The proteins are named B, $D_1$, $D_2$, $D_3$, E, F and G. An eighth protein, B' has been found in HeLa cells which turned out to be an alternatively spliced form of protein B, differing only in the 11 C–terminal residues (Chu & Elkon, 1991; van Dam *et al.*, 1989). The sequence alignment of the known Sm and Sm–like proteins revealed two conserved sequence motifs, named *Sm1* and *Sm2* (Séraphin, 1995; Hermann *et al.*, 1995; Cooper *et al.*, 1995).

The human Sm proteins assemble into sub–complexes in the absence of snRNA. The complexes formed are $D_1D_2$, $D_3B$ ($D_3B'$ as well) and EFG (Lehmeier *et al.*, 1994; Hermann *et al.*, 1995; Raker *et al.*, 1996). The importance of the Sm motifs has been demonstrated for the $D_3B$ complex: the presence of the Sm motifs was necessary and sufficient for sub–complex formation (Herman *et al.*, 1995). The snRNA binding site of the core Sm protein complex was identified almost twenty years ago. Branlant *et al.* identified a conserved short uridine rich single stranded sequence with the general structure Pu–A–$(U)_{4-6}$–G–Pu in the snRNAs U1, U2, U4 and U5 as the binding sites of Sm proteins (Branlant *et al.*, 1982). The efficiency of binding to this Sm site is modulated

by snRNP specific proteins and neighbouring RNA sequences (Jarmolowski & Mattaj, 1993; Nelissen *et al.*, 1994). Recently published binding studies with nonamer oligonucleotides have shown that the Sm site alone is sufficient to form the complete core domain. In addition to the uridine bases, the second adenosine and the 2'–hydroxyl groups of the ribose moieties turned out to be essential for binding (Raker *et al.*, 1999).

In the absence of the snRNA (the Sm site) three stable sub–complexes are formed between Sm proteins. They associate in an orderly manner with the Sm site of the snRNA: EFG, together with $D_1D_2$, binds first forming a stable subcore, then $D_3B$ joins the complex completing the assembly. It has been shown that none of the sub–complexes can bind to the snRNA in a stable manner, an essential prerequisite for snRNP formation (Fischer *et al.*, 1985, Feeney *et al.*, 1989; Raker *et al.*, 1996). The pairwise interaction between the individual Sm proteins in the core complex was studied by the application of the yeast two–hybrid system. The information provided by these experiments was essential to construct the first model of human core Sm complex (Kambach *et al.*, 1999; Kambach & Nagai, 1999). After acquiring an $N^7$–monomethylguanosine cap ($m^7G$) during transcription, the snRNA is transported to the cytoplasm where the assembly with the Sm core domain takes place (Mattaj & De Robertis, 1985). The complete core domain is necessary for the hypermethylation of the $m^7G$ cap to a 2,2,7–trimethyl–guanosine ($m_3G$) cap (Mattaj, 1986). The nuclear import of the spliceosomal snRNPs depends on a bipartite signal consisting of the $m_3G$ cap and the complete core domain (Hamm *et al.*, 1990; Fischer *et al.*, 1993; Plessel *et al.*, 1994), whereas for the U4 and U5 snRNPs the presence of the core domain is sufficient (Palacios *et al.*, 1997).

Recently, the crystal structure of two core sub–complexes, $D_1D_2$ and $D_3B$ have been determined by X–ray crystallography (Kambach *et al.*, 1999). The four Sm proteins show a common fold containing a short N–terminal α–helix followed by a strongly bent, five–stranded, antiparallel β–sheet. The *Sm1* motif is formed by strands 1–3 of the β–

sheet, whereas *Sm2* motif is constituted by strands 4 and 5. The *Sm1* and *Sm2* motifs are connected by a loop of variable length. The two Sm proteins in the dimeric complexes interact via their 4$^{th}$ and 5$^{th}$ β–strands forming a continuous inter–subunit β–sheet.

Based on the core sub–complex structures Kambach *et al.* proposed a model for the entire Sm core domain, that is consistent with all the structural (EM & X–ray), genetic and biochemical data available up–to–date. The model consists of seven different Sm proteins arranged in a heptameric ring, interfacing each other the same way as in the sub–complex dimers. The order of the individual Sm proteins in the ring is consistent with the experimentally observed pairwise interactions within the core Sm domain (Fury *et al.*, 1997; Camasses *et al.*, 1997). The inner surface of the ring is lined with positively charged residues and the size of the central hole (~20 Å diameter), if considering only main chain atoms, can easily accommodate single stranded RNA. These facts strongly suggests that the snRNA binds in the central hole, an assumption which is supported by UV cross–linking experiments, where the AUU sequence of the Sm site was cross–linked to Sm protein G (Heinrichs *et al.*, 1992).

## 4.3 Archaeal Sm–like proteins

There are more and more fully sequenced archaebacterial genomes available which allows to search for and identify the archaeal counterparts of eukaryotic proteins. Sensitive searches in sequence databases revealed the existence of some archaeal proteins related to Sm and Sm–like proteins (Salgado–Garido *et al.*, 1999). *Methanobacterium thermoautrotrophicum* (Smith *et al.*, 1997), *Aeropyrum pernix* K1 (Kawarabayasi *et al.*, 1999) and *Archaeoglobus fulgidus* (Klenk *et al.*, 1997) were found to encode two Sm–like proteins, while *Pyrococcus horikoshii* (Kawarabayasi *et al.*, 1998) and *Pyrococcus abyssi* (Poch *et al.*, submitted) encode only a single Sm–related protein. The finding of Sm–

related proteins in various archaebacterial genomes suggests that an Sm–like protein was present in the latest common ancestor of archaeons and eukaryotes, and that the diverse eukaryotic Sm and Sm–like proteins originate from a single precursor (Salgado–Garido *et al.*, 1999).

The crystal structure of one of the Sm–related proteins of *Archaeoglobus fulgidus* has been solved and refined as part of this thesis. The Sm–like protein, named AF–Sm2 forms hexameric rings both in the crystal structure and in solution, shown by gel filtration experiments in the latter case. The nature of the major interactions between AF–Sm2 monomers is essentially the same as what has been found in the human Sm core domain sub–complexes (Kambach *et al.*, 1999). The central hole of the ring is smaller than that of the heptameric complex proposed by Kambach *et al.*, but appears still to be large enough to accommodate single–stranded RNA. The function of the archaeal Sm–like proteins is currently unknown.

## 4.4 References

Achsel, T., Brahms, H., Kastner, B., Bachi, A., Wilm, M. & Lührmann, R. (1999). A doughnut–shaped heteromer of human Sm–like proteins binds to the 3'–end of U6 snRNA, thereby facilitating U4/U6 duplex formation *in vitro*. *EMBO J.*, **18**, 5789–5802.

Baserga, S.J. & Steitz, J.A. (1993). The diverse world of small ribonucleoproteins. In *The RNA World*, Gesteland, R.F. & Atkins, J.F., eds. (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press), pp. 359–381.

Beggs, J.D. (1995). Yeast splicing factors and genetics strategies for their analysis. In *Pre–mRNA Processing*, Edited by Lamond, A.I.. (Austin, TX, RG Landes Co), pp. 79–95.

Bordonné, R. & Tarassov, I. (1996). The Yeast SME1 gene encodes the homologue of the human E core protein. *Gene*, **176**, 111–117.

Branlant. C., Krol, A., Ebel, J.–P., Lazar, E., Haendler, B. & Jacob, M. (1982). U2 RNA shares a structural domain with U1, U4 and U5 RNAs. *EMBO J.* **1**, 1259–1265.

Burge, C.B., Tuschl, T. & Sharp, P.A. (1999). Splicing of precursors to mRNAs by the spliceosomes. In *The RNA World*, 2nd edn., Gesteland, R.F., Chech, T.R. & Atkins, J.F., eds. (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press), pp. 525–560.

Camasses, A., Bragado, N.E., Martin, R., Séraphin, B. & Bordonné, R. (1998). Interactions within the yeast Sm core complex: from proteins to amino acids. *Mol. Cell. Biol.* **18**, 1956–1966.

Caspary, F & Séraphin, B. (1998). The yeast U2A'/U2B'' complex is required for pre–spliceosome formation. *EMBO J.* **17**, 6348–6358.

Chu, J.–L. & Elkon, K.B. (1991). The small nuclear ribonucleoproteins, SmB and B', are products of a single gene. *Gene* **97**, 311–312.

Cooper, M., Johnston, L.H. & Beggs, J.D. (1995). Identification and characterisation of Uss1p (Sdb23p): a novel U6 snRNA–associated protein with significant similarity to core proteins of small nuclear ribonucleoproteins. *EMBO J.* **14**, 2066–2075.

Feeney, R.J., Suaterer, R.A., Feeney, J.L. & Zieve, G.W. (1989). Cytoplasmic assembly

and nuclear accumulation of mature small nuclear ribonucleoprotein particles. *J. Biol. Chem.* **264**, 5776–5783.

Fischer, D.E., Conner, G.E., Reeves, W.H., Wisniewolski, R. & Blobel, G. (1985). Small nuclear ribonucleoprotein particle assembly in vivo: demonstration of a 6S RNA–free core precursor and posttranslational modification. *Cell* **42**, 751–758.

Fischer, U., Sumpter, V., Sekine, M., Satoh, T. & Lührmann, R. (1993). Nucleo–cytoplasmic transport of U snRNPs: definition of a nuclear location signal in the Sm core domain that binds a transport receptor independently of the $m_3G$ cap. *EMBO J.* **12**, 573–583.

Fromont, R.M., Rain, J.C. & Legrain, P. (1997). Toward a functional analysis of yeast genome through exhaustive two–hybrid screens. *Nature Genet.* **16**, 277–282.

Fury, M.G., Zhang, W., Christodoulopoulos, I. & Zieve, G.W. (1997). Multiple protein:protein interactions between the snRNP common core proteins. *Exp. Cell. Res.* **237**, 63–69.

Hamm, J., Darzynkiewicz, E., Tahara, S.M. & Mattaj, I.W. (1990). The trimethylguanosine cap structure of U1 snRNA is a component of a bipartite nuclear targeting signal. *Cell* **62**, 569–577.

Heinrichs, V., Bach, M., Winkelman, G. & Lührmann, R. (1990). U1–specific protein C needed for efficient complex formation of U1 snRNP with 5' splice site. *Science* **247**, 69–72.

Heinrichs, V., Hackl, W. & Lührmann, R. (1992). Direct binding of small nuclear ribonucleoprotein G to the Sm site of small nuclear RNA. Ultraviolet light cross–linking of protein G to the AUU stretch within the Sm site (AAUUUGUGG) of U1 small nuclear ribonucleoprotein reconstituted in vitro. *J. Mol. Biol.* **227**, 15–28.

Hermann, H., Fabrizio, P., Raker, V.A., Foulaki, K., Hornig, H., Brahms, H. & Lührmann, R. (1995). snRNP Sm proteins share two evolutionary conserved sequence motifs which are involved in Sm protein–protein interaction. *EMBO J.* **14**, 2076–2088.

Jarmolowski, A. & Mattaj, I.W. (1993). The determinants for Sm protein binding to *Xenopus* U1 and U5 snRNAs are complex and non–identical. *EMBO J.* **12**, 223–232.

Kambach, C., Walke, S., Young, R., Avis, J.M., de la Fortelle, E., Raker, V.A., Lührmann, R., Li, J. & Nagai, K. (1999). Crystal structure of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs. *Cell* **96**, 375–387.

Kambach, C., Walke, S. & Nagai, K., (1999). Structure and assembly of the spliceosomal small nuclear ribonucleoprotein particles. *Curr. Opin. Struct. Biol.* **9**, 222–230.

Kastner, B., Kronstädt, U., Bach, M. & Lührmann, R. (1992). Structure of the small nuclear RNP particle U1: identification of the two structural protuberances with RNP–antigens A and 70K. *J. Cell. Biol.* **116**, 839–849.

Kastner, B., Bach, M. & Lührmann, R. (1990). Electron microscopy of small nuclear ribonucleoprotein (snRNP) particles U2 and U5: evidence for a common structure–determining principle in the major U snRNP family. *Proc. Natl. Acad. Sci.* **87**, 1710–1714.

Kastner, B., Bach, M. & Lührmann, R. (1990). Electron microscopy of snRNPs U2. U4/U6 and U5: evidence for a common structure–determining principle in the major U snRNP family. *Mol. Biol. Rep.* **14**, 171.

Kastner, B., Bach, M. & Lührmann, R. (1991). Electron microscopy of U4/U6 snRNP reveals a Y–shaped U4 and U6 RNA containing domain protruding from the U4 core RNP. *J. Cell. Biol.* **112**, 1065–1072.

Kastner, B. (1998). Purification and elecron microscopy of spliceosomal snRNPs. In *RNP Particles, Splicing and Autoimmune Disease*. Edited by Schenkel, J., Berlin, Springer Verlag, pp. 95–140.

Kawarabayasi, Y. *et al.* (1998). Complete sequence and gene organization of the genome of a hyper–thermophylic achaebacterium, *Pyrococcus horikoshii* OT3. *DNA Res.* **5**, 55–76.

Kawarabayasi, Y. *et al.* (1999). Complete sequence of an aerobic hyper–thermophylic crenachaeon, Aeropyrum pernix K1. *DNA Res.* **6**, 83–101.

Klenk, H.–P. *et al.* (1997). The complete genome sequence of the hyperthermophilic, sulphate–reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390**, 364–370.

Krämer, A. (1996). The structure and function of proteins involved in mammalian pre–mRNA splicing. *Annu. Rev. Biochem.* **65**, 367–409.

Lehmeier, T., Raker, V.A., Hermann, H & Lührmann, R. (1994). cDNA cloning of the Sm proteins $D_2$ and $D_3$ from human small nuclear ribonucleoproteins: evidence for a direct $D_1$ $-D_2$ interaction. *Proc. Natl. Acad. Sci. USA* **91**, 12317–12321.

Lerner, M.R. & Steitz, J.A. (1979). Antibodies to small nuclear RNAs complexed with proteins are produced by patients with systemic lupus erythmatosus. *Proc. Natl. Acad. Sci. USA* **76**, 5495–5499.

Liautard, J.–P., Sri–Widada, J., Brunel, C. & Jeanteur, P. (1982). Structural organization of ribonucleoproteins containing small nuclear RNAs from HeLa cells. *J. Mol. Biol.* **162**, 623–643.

Lührmann, R., Kastner, B. & Bach, M. (1990). Structure of spliceosomal snRNPs and their role in pre–mRNA splicing. *Biochim. Biophys. Acta*, **1087**, 265–292.

Madhani, H.D. & Guthrie, C. (1994). Dynamic RNA–RNA interactions in the spliceosome. *Annu. Rev. Genet.* **28**, 1–26.

Mattaj, I.W. & De Robertis, E.M. (1985). Nuclear segregation of U2 snRNA requires binding of specific snRNP proteins. *Cell* **40**, 111–118.

Mattaj, I.W. (1986). Cap trimethylation of U snRNA is cytoplasmic and dependent on U snRNP protein binding. *Cell* **46**, 905–911.

Mattaj, I.W., Tollervey, D. and Séraphin, B. (1993). Small nuclear RNAs in messenger RNA and ribosomal RNA processing. *FASEB J.* **7**, 47–53.

Moore, M.J., Query, C.C. & Sharp, P.A. (1993). Splicing of precursors to messenger RNAs by the spliceosome. In *The RNA World*, Gesteland, R.F. & Atkins, J.F., eds. (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press), pp. 303–357.

Müller, S., Wolpensinger, B., Angenitzki, M., Engel, A., Sperling, J. & Sperling, R. (1998). A supraspliceosome model for large nuclear ribonucleoprotein particles based on mass determinations by scanning transmission electron microscopy. *J. Mol. Biol.* **283**, 383–394.

Nagai, K. & Mattaj, I.W. (1994). RNA–protein interactions in the splicing snRNPs. In *RNA–Protein interactions.* Nagai, K. and Mattaj, I.W. eds. (Oxford, Oxford University Press), pp. 150–177.

Nagai, K., Oubridge, C., Jessen, T.H., Li, J. & Evans, P.R. (1990). Crystal structure of the RNA–binding domain of the U1 small nuclear ribonucleoprotein A. *Nature* **348**, 515–520.

Nelissen, R.L., Will, C.L., van Venrooij, W.J. & Lührmann, R. (1994). The assoxiation of the U1–specific 70K and C proteins with U1 snRNPs is mediated in part by common U snRNP proteins. *EMBO J.* **13**, 4113–4125.

Nelissen, R.L.H., Heinrichs, V., Habets, W.J., Simons, F., Lührmann, R. & van Venrooij, W.J. (1991). Zinc finger–like structure in U1–specific protein C is essential for specific binding to U1 snRNP. *Nucleic Acids Res.* **19**, 449–454.

Newman, A.J. & Norman, C. (1992). U5 snRNA interacts with exon sequences at 5' and 3' splice sites. *Cell* **68**, 743–754.

O'Keefe, R.T., Norman, C, & Newman, A.J. (1996). The invariant U5 snRNA loop 1 sequence is dispensable for the first catalytic step of splicing in yeast. *Cell* **86**, 679–689.

Oubridge, C., Ito, N., Evans, P.R., Teo, H.C. & Nagai, K. (1994). Crystal structure at 1.92 Å resolution of the RNA binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. *Nature* **372**, 432–438.

Palacios, I., Hetzer, M., Adam, S.A. & Mattaj, I.W. (1997). Nuclear import of U snRNPs requires importin *b*. *EMBO J.* **16**, 6783–6792.

Plessel, G., Fischer, U. & Lührmann, R. (1994). $m_3G$ cap hypermethylation of U1 small nuclear ribonucleoprotein (snRNP) in vitro: evidence that the U1 small nuclear RNA–(guanosine–N2)–methyl–transferase is a non–snRNP cytoplasmic protein that requires a binding site on the Sm core domain. *Mol. Cell. Biol.* **14**, 4160–4172.

Poch *et al.*, (2000). Multiple comparison of Pyrococcus species: evidence for gene transfer between bacteria and archaea. Submitted. in Genome Research.

Price, S.R., Evans, P.R. & Nagai, K. (1998). Crystal structure of the spliceosomal U2B''–U2A' protein complex bound to a fragment of U2 small nuclear RNA. *Nature* **394**, 645–650.

Query, C.C., Bentley, R.C. & Keene, J.D. (1989). A common RNA recognition motif identified within a defined U1 RNA binding domain of the 70K U1 snRNP protein. *Cell* **57**, 89–101.

Raker, V.A., Plessel, G. & Lührmann, R. (1996). The snRNP core assembly pathway: identification of stable core protein heteromeric complexes and an snRNP subcore particle *in vitro*. *EMBO J.* **15**, 2256–2269.

Raker, V.A., Hartmuth, K., Kastner, B. & Lührmann, R. (1999). Spliceosomal U snRNP core assembly: Sm proteins assemble onto an Sm site RNA nonanucleotide in a specific and thermodynamically stable manner. *Mol. Cell. Biol.* **19**, 6554–6565.

Roy, J., Zheng, B., Rymond, B.C. & Woolford, J. (1995). Structurally related but functionally distinct yeast Sm D core small nuclear ribonucleoprotein particle proteins. *Mol. Cell. Biol.* **15**, 445–455.

Rymond, B.C. (1993). Convergent transcripts of the yeast PRP38–SMD1 locus encode two essential splicing factors, including the D1 core polypeptide of small nuclear ribonucleoprotein particles. *Proc. Natl. Acad. Sci. USA* **90**, 848–852.

Salgado–Garrido, J., Bragado–Nilsson, E., Kandels–Lewis, S. & Séraphin, B. (1999). Sm and Sm–like proteins assemble in two related complexes of deep evolutionary origin. *EMBO J.* **18**, 3451–3462.

Scherly, D., Boelens, W., van Venrooij, W.J., Dathan, N.A., Hamm, J. & Mattaj, I.W. (1989). Identification of the RNA binding segment of human U1A protein and definition of its binding site on U1 snRNA. *EMBO J.* **8**, 4163–4170.

Scherly, D., Boelens, W., Dathan, N.A., van Venrooij, W.J. & Mattaj, I.W. (1990). Major determinants of the specificity of interaction between small nuclear ribonucleoprotein–

U2B'' and their cognate RNAs. *Nature* **345**, 502–506.

Séraphin, B. (1995). Sm and Sm–like proteins belong to a large family: identification of proteins of the U6 as well as the U1, U2, U4 and U5 snRNPs. *EMBO J.* **14**, 2089–2098.

Seto, A.G., Zaug, A.J., Sobel, S.G., Wolin, S.L. & Cech, T.R. (1999). Saccharomyces cerevisiae telomerase is an Sm small nuclear ribonucleoprotein particle. *Nature* **401**, 177–180.

Smith, D.R. *et al.* (1997). Complete genome sequence of *Methanobacterium thermoautrotrophicum* ΔH: functional analysis and comparative genomics. *J. Bacteriol.* **179**, 7135–7155.

Smith, H.O., Tabiti, K., Schaffner, G., Soldati, D., Albrecht, U., & Birnstiel, M.L. (1991). Two–step affinity purification of U7 small nuclear ribonucleoprotein particles using complementary biotinylated 2'–*O*–methyl oligoribonucleotides. *Proc. Natl. Acad. Sci. USA* **88**, 9784–9788.

Sontheimer, E.J. & Steitz, J.A. (1993). The U5 and U6 small nuclear RNAs as active site components of the spliceosome. *Science* **262**, 1989–1996.

Staley, J.P. & Guthrie, C. (1998). Mechanical devices of the spliceosome: motors, clocks, springs and things. *Cell* **92**, 315–326.

van Dam, A., Winkel, I., Zijlstra–Baalbergen, J., Smeenk, R. & Cuypers, H.T. (1989). Cloned human snRNP proteins B and B' differ only in their carboxy–terminal part. *EMBO*

*J.* **8**, 3853–3860.

Will, C.L. & Lührmann, R. (1997). Protein functions in pre–mRNA splicing. *Curr. Opin. Cell. Biol.* **9**, 320–328.

# Chapter 5

# Structure determination of AF–Sm2 from *Archaeoglobus fulgidus*

## 5.1 Sample preparation: cloning, expression and purification

The complete genome of *Archaeoglobus fulgidus* has been sequenced and deposited in sequence databanks (Klenk *et al.*, 1997). Sensitive searches in the databanks revealed two ORFs encoding two Sm–related proteins of 77 and 75 amino acid size (Salgado–Garido *et al.*, 1999). Although both proteins have been cloned from genomic DNA of *Archaeoglobus fulgidus*, the discussion will be however restricted to the 75 amino acid AF–Sm2 protein, for which an X–ray structure has been determined.

### 5.1.1 The cloning of the gene encoding AF–Sm2

The genomic DNA of *Archaeoglobus fulgidus* were obtained from the German Collection of Micro–organisms, Braunschweig, Germany. The database identifier of the AF–Sm2 gene is AF0362. Based on this sequence oligonucleotides have been designed in order to amplify the gene and to incorporate NcoI and KpnI restriction sites for further cloning steps (Figure 5.1).

**Oliogo Design: SM2-like A.fulgidus**

**N-Terminus**

```
 M  V  L
atggtgcttccaaatcagatggtaaagtcaatggtgggaaagataataaggg
taccacgaaggtttagtctaccatttcagttaccacccttctattattccc
```

**Oligo Name: SM2F**

NcoI: C/CATGG

```
          M  V
5-ATAATTCCATGGTGCTTCCAAATCAGATGGTAAAGTCAATGGTGG
```

**C-Terminus**

```
       TaiI            BsoFI
       MaeII           CviJI   MboII
       |               |||     |
ggtaataacgtcgttctaatccagccgcaagaagaatga  228
ccattattgcagcaagattaggtcggcgttcttcttact
       |               |||     |
       197             212     220
```

**Oligo Name: SM2B**

KpnI=GGTACC

5'-ATAATAGGTACCTCATTCTTCTTGCGGCTGGATTAGAACGACGTTATTACC

**Figure 5.1** The oligonucleotides designed for the PCR amplification of the *A. fulgidus* gene AF0362 encoding AF–Sm2. The PCR product incorporates NcoI and KpnI restriction sites.

After the initial heating step (95° for 5 min) the following PCR protocol was used in 30 cycles: 55°(1 min)⟶72°(0.5 min)⟶94°(1 min). The composition of the PCR reaction mix was as follows:

- 200 µM dNTP (5 µl of 2 mM stock)
- 1x AmpliTaq buffer from Perkin–Elmer
- 50 pmol of primer oligonucleotides (~2.5 µl each)
- 1 µl of genomic DNA
- 1 µl of Taq polymerase Perkin–Elmer

The total reaction mixture was topped up to 50 µl volume with sterile water. Following the PCR amplification the *PCR Cleanup* kit from Promega was used to purify the amplified AF–Sm2 gene. The purified PCR product was eluted with 50 µl TE buffer (10 mM Tris, 1 mM EDTA, pH 8.0). 10 µl of this eluate was digested with 2 units of NcoI and KpnI restriction enzymes in a total volume of 50 µl using the Yellow buffer. The enzymes and the buffer were purchased from Fermentas, Lithuania. The product of the reaction was purified with the *DNA Cleanup* kit from Promega, and eluted in 50 µl TE buffer.

AF–Sm2 was expressed as a GST fusion protein. The expression vector used, pET24H6·GST·TEV, is a modified pET24 vector (Studier *et al.*, 1990) with an N–terminal His–tag, a GST–fragment and a TEV protease cleavage site upstream from the NcoI restriction site (Figure 5.2 and 5.3).

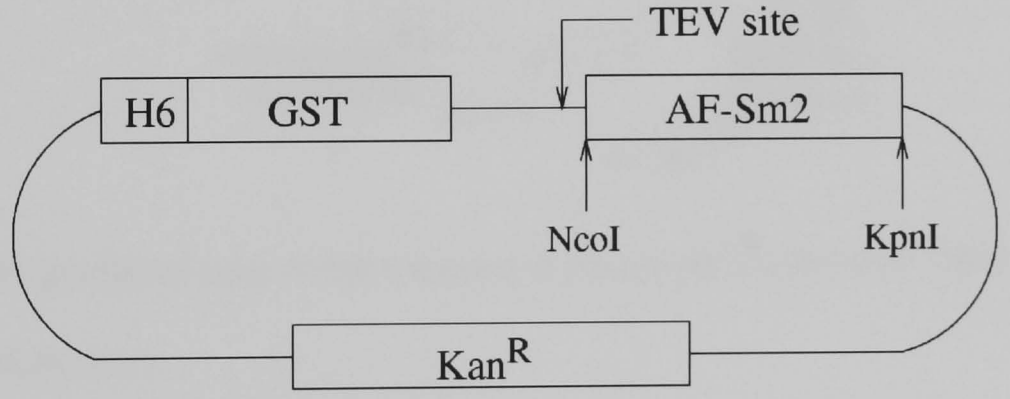



**Figure 5.2** Schematic drawing of the expression vector with the GST–AF–Sm2 construct. The expression cassette of the fusion protein contains an N–terminal His–tag, a GST fragment and a TEV protease cleavage site 5' of the NcoI restriction site (Figure 5.4).

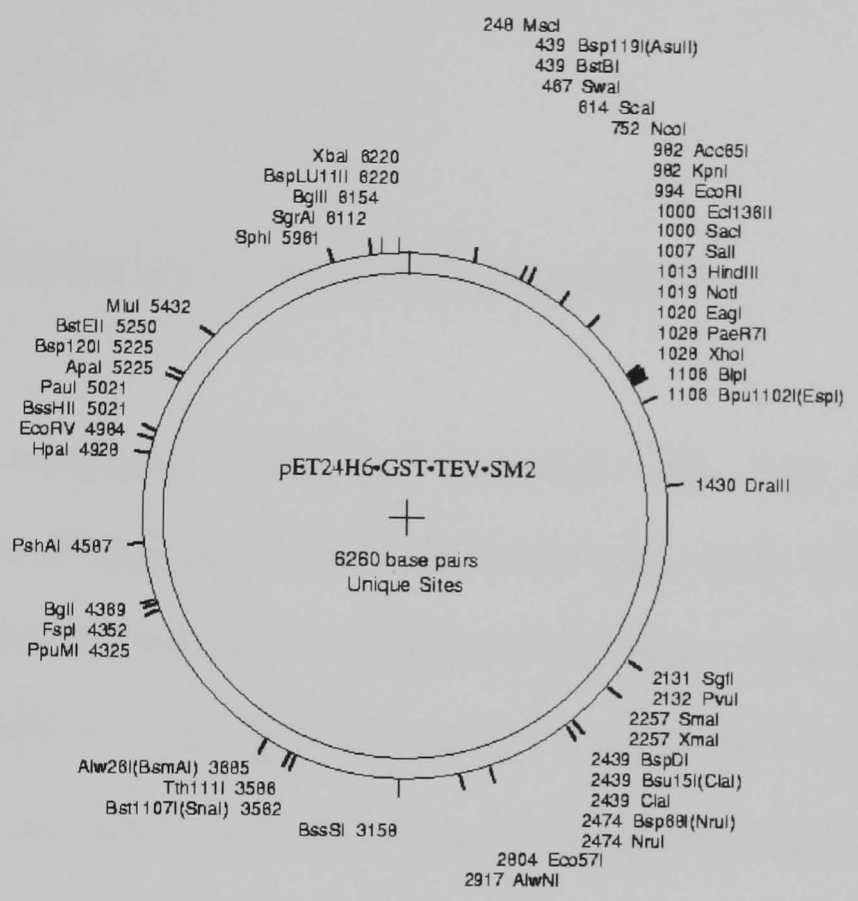



**Figure 5.3** The graphical map of the expression vector pET24H6·GST·TEV·AF–Sm2 with unique restriction sites.

Prior to ligation the vector was cleaved with NcoI and KpnI, followed by the direct addition of phosphatase and phospatase buffer in order to remove the terminal phosphate groups left by the restriction enzymes. The cleaved vector was then purified the same way as it was described for the AF–Sm2 gene. The ligation reaction mixture was the following:

- 1 µl (~5 ng) cleaved and purified expression vector
- 3 µl NcoI–KpnI cleaved and purified AF–Sm2 fragment
- 1 µl 10X ligation buffer
- 1 µl T4 ligase

The total volume was 10 µl. 1 µl of the reaction mixture was used to transform DH5α competent cells, which were plated on kanamycin containing agarose plates. 16–20 hours later two colonies were picked and cell cultures were grown in order to obtain expression plasmids. The plasmid preparation was purified with a plasmid purification kit from Qiagen, Hamburg. The two plasmid preps were checked with restriction enzymes whether they contained the whole construct (Figure 5.4).

### 5.1.2 Expression of the GST–AF–Sm2 fusion protein

For protein expression BL21(DE3) competent cells from Novagen were transformed with the expression vector tested for the presence of the whole GST–fusion (Figure 5.4) and plated on kanamycin containing agarose plates. After 16–20 hours the colonies were suspended in a few ml of LB medium to inoculate 8 times half litre LB medium containing 200 µg/ml kanamycin. The cell cultures were induced with 0.4 mM IPTG when their O.D. measured at 595 nm reached 0.7. After 2–3 hours of induction the cell cultures were centrifuged in one litre flasks fitting into a swinging bucket rotor, resuspended in fresh LB and collected in a single storage vial. The cell pellet was frozen in liquid nitrogen for storage or used immediately. The cell pellet was suspended and lysed in the following lysis buffer (40 ml):

- 50 mM TRIS, pH 8.0
- 100 µg/ml lysozyme
- 1 mg Bovine DNase I
- 12 mM $MgCl_2$
- 10 mM imidazole
- 1 tablet of EDTA free protease inhibitor cocktail *Complete* from Boehringer–Mannheim

**pET24H6-GST-TEV-SM2 [1 to 981] -> 1-phase Translation**

Translation of Expression Casette (His-GST-TEV-SM)

```
1/1                                                 31/11
ATG aaa cat cac cat cac cat cac aac act agt agc aat tcc atg tcc cct ata cta ggt
M   K   H   H   H   H   H   H   N   T   S   S   N   S   M   S   P   I   L   G
61/21                                               91/31
tat tgg aaa att aag ggc ctt gtg caa ccc act cga ctt ctt ttg gaa tat ctt gaa gaa
Y   W   K   I   K   G   L   V   Q   P   T   R   L   L   L   E   Y   L   E   E
121/41                                              151/51
aaa tat gaa gag cat ttg tat gag cgc gat gaa ggt gat aaa tgg cga aac aaa aag ttt
K   Y   E   E   H   L   Y   E   R   D   E   G   D   K   W   R   N   K   K   F
181/61                                              211/71
gaa ttg ggt ttg gag ttt ccc aat ctt cct tat tat att gat ggt gat gtt aaa tta aca
E   L   G   L   E   F   P   N   L   P   Y   Y   I   D   G   D   V   K   L   T
241/81                                              271/91
cag tct atg gcc atc ata cgt tat ata gct gac aag cac aac atg ttg ggt ggt tgt cca
Q   S   M   A   I   I   R   Y   I   A   D   K   H   N   M   L   G   G   C   P
301/101                                             331/111
aaa gag cgt gca gag att tca atg ctt gaa gga gcg gtt ttg gat att aga tac ggt gtt
K   E   R   A   E   I   S   M   L   E   G   A   V   L   D   I   R   Y   G   V
361/121                                             391/131
tcg aga att gca tat agt aaa gac ttt gaa act ctc aaa gtt gat ttt ctt agc aag cta
S   R   I   A   Y   S   K   D   F   E   T   L   K   V   D   F   L   S   K   L
421/141                                             451/151
cct gaa atg ctg aaa atg ttc gaa gat cgt tta tgt cat aaa aca tat tta aat ggt gat
P   E   M   L   K   M   F   E   D   R   L   C   H   K   T   Y   L   N   G   D
481/161                                             511/171
cat gta acc cat cct gac ttc atg ttg tat gac gct ctt gat gtt gtt tta tac atg gac
H   V   T   H   P   D   F   M   L   Y   D   A   L   D   V   V   L   Y   M   D
541/181                                             571/191
cca atg tgc ctg gat gcg ttc cca aaa tta gtt tgt ttt aaa aaa cgt att gaa gct atc
P   M   C   L   D   A   F   P   K   L   V   C   F   K   K   R   I   E   A   I
601/201                                             631/211
cca caa att gat aag tac ttg aaa tcc agc aag tat ata gca tgg cct ttg cag ggc tgg
P   Q   I   D   K   Y   L   K   S   S   K   Y   I   A   W   P   L   Q   G   W
661/221                                             691/231
caa gcc acg ttt ggt ggt ggc gac cat cct cca act agt gga tct ggt ggt ggt ggc gga
Q   A   T   F   G   G   G   D   H   P   P   T   S   G   S   G   G   G   G   G
721/241                                             751/251
tcc atg agc gag aat ctt tat ttt cag ggc gCC ATG Gtg ctt cca aat cag atg gta aag
S   M   S   E   N   L   Y   F   Q   G   A   M   V   L   P   N   Q   M   V   K
781/261                                             811/271
tca atg gtg gga aag ata ata agg gtc gaa atg aag ggc gag gag aac cag tta gtc ggg
S   M   V   G   K   I   I   R   V   E   M   K   G   E   E   N   Q   L   V   G
841/281                                             871/291
aaa ctt gag ggt gtg gac gac tac atg aac cta tac ttg aca aac gcg atg gag tgc aag
K   L   E   G   V   D   D   Y   M   N   L   Y   L   T   N   A   M   E   C   K
901/301                                             931/311
ggg gag gag aag gta agg agc ctg gga gaa ata gtg ctc aga ggt aat aac gtc gtt cta
G   E   E   K   V   R   S   L   G   E   I   V   L   R   G   N   N   V   V   L
961/321
atc cag ccg caa gaa gaa tga
I   Q   P   Q   E   E   *
```

**Figure 5.4** The expression cassette of the whole GST–fusion. The linker with a TEV cleavage site is shown in blue, the AF–Sm2 sequence is red. One has to note that the two last residues of the linker (G A) are not removed by TEV protease, thus they form an N–terminal extension not present in the wild type protein increasing its length to 77 amino acids.

The lysed cells were finally passed through a French press to maximise cell lysis and ultracentrifuged at 45000 rpm to get rid of cell debris. Both the pellet and the supernatant were analysed by gel electrophoresis and both were shown to contain the fusion protein,

although more than the half, approximately 60% of the total protein was present in the supernatant.

### 5.1.3 Purification

The supernatant was purified in two passes. Half of it was loaded onto a 10 ml Ni–agarose column equilibrated with 20 mM Tris–HCl, pH 8.0, 150 mM NaCl and 10 mM imidazole and washed with two column volumes of the same buffer. After washing, the fusion protein was eluted with the same buffer but containing 250 mM imidazole this time. After analysis by SDS–PAGE the best fractions were collected (~20 ml) in a dialysis tube. Before closing the dialysis tube 300 µl crude TEV protease preparation was added to it. The dialysis was carried out at 4°C against 20 mM Tris–HCl, pH 8.0, 150 mM NaCl. After two days of dialysis the protein preparation was placed into an 86°C water bath for 15 minutes. The heating step essentially removed all the proteins but AF–Sm2 (Figure 5.5). The precipitated protein was removed by centrifugation, and the Sm protein was concentrated to ~11 mg/ml. AF–Sm2 contains only two tyrosine residues therefore it has a low molar extinction coefficient at 280 nm. Therefore, the concentration of the protein was measured by the Lowry method (Lowry *et al.*, 1951.)

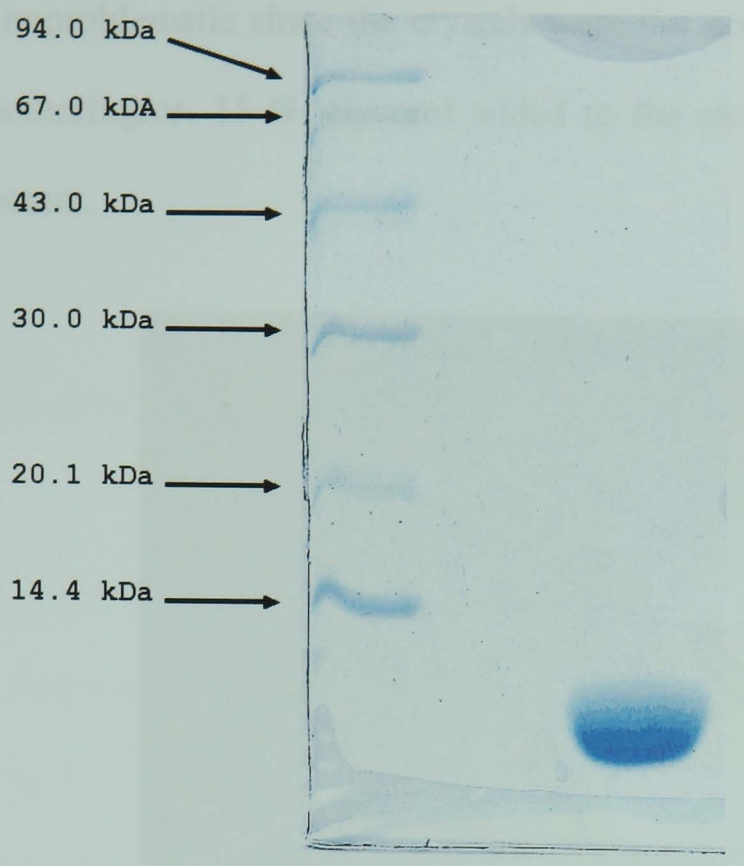



**Figure 5.5** SDS–PAGE of purified and concentrated AF–Sm2 with molecular weight markers.

## 5.2 Crystallisation, data collection and processing

### 5.2.1 Crystallisation of AF–Sm2

Crystallisation setups were done using the vapour diffusion method with hanging drops. A 48 condition in–house sparse matrix screen (Zeelen *et al.*, 1994) was tried with immediate success. Long hexagonal needles were obtained from a low pH condition containing ammonium and lithium sulphate as precipitants. The crystallisation condition was optimised resulting in a more favourable shape of the crystals: they became shorter (0.15–0.2 mm) but their diameter increased dramatically (Figure 5.6). The optimised crystallisation condition was the following: 2.5 M ammonium sulphate, 200 mM lithium

sulphate, 120 mM sodium acetate buffer, pH 3.6. Searching for a suitable freezing condition was unproblematic since the crystals were not sensitive at all to changes in the surrounding motherliquor. 15 % glycerol added to the motherliquor was found to be a good cryoprotectant.

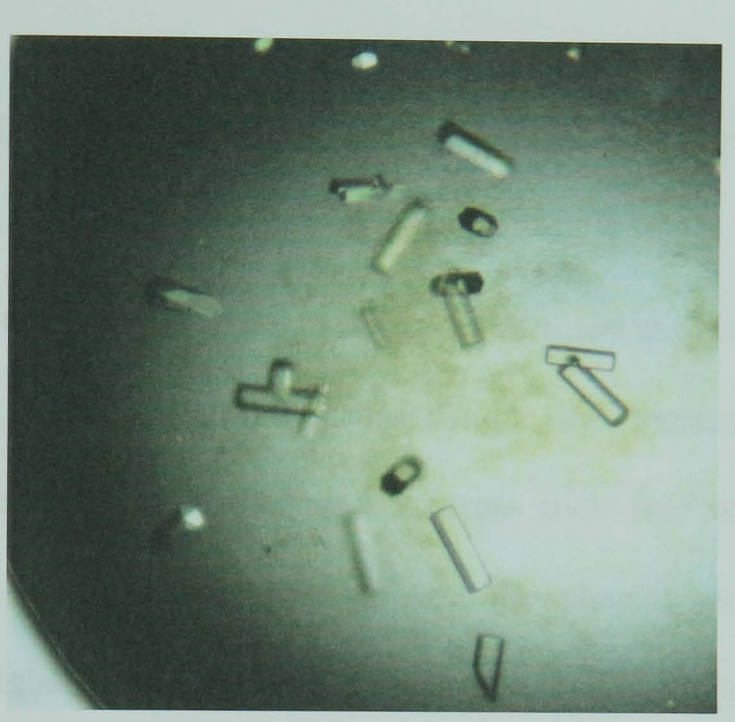


**Figure 5.6** Hexagonal AF–Sm2 crystals.

### 5.2.2 The heavy atom derivative

At the time of structure determination there were no Sm or Sm–like protein structures deposited in the PDB, therefore we had to look for (a) heavy atom derivative(s) to obtain initial phase information. An alternative to MIR phasing was to prepare Se–Met protein and collect MAD data at a synchrotron site. This seemed to be possible, since there were six methionine residues in an AF–Sm2 molecule. However, soaking existing crystals with solutions of heavy atom compounds is more straightforward than preparing Se–Met derivative crystals and was attempted first. Mercury compounds seemed to be the most suitable, because AF–Sm2 has a single cysteine residue capable of binding mercury with high affinity.

Crystals were soaked overnight with their mother liquor containing additional 15%

glycerol and 5 mM methyl–mercury acetate (MMA). No macroscopic change in the crystals were seen as a result of MMA–soaking. The first soak was tested in the X–ray beam and a complete data set was collected. The data collection and data quality will be discussed later in this chapter.

### 5.2.3 The collection and processing of the native and derivative data

The first native and derivative data sets were collected at the home source using an Enraf–Nonius rotating anode generator. Interestingly the mercury derivative soaked crystals diffracted better than the native ones whereas their size was approximately the same. The native data set was only used to test the phasing power of the mercury derivative, and then a higher resolution data set was collected at a synchrotron beamline.

The native data set used for refinement was collected at the BM14 beamline of ESRF on a MarCCD detector. The crystals were transported already frozen in a Dewar–container filled with liquid nitrogen. 15% glycerol was added to the mother liquor as cryoprotectant before flash–freezing the crystals. During data collection the crystal was kept in a dry nitrogen gas stream of 100 K in order to prevent radiation damage. In total 45 degrees of oscillation data were collected in a single pass with 1 degree oscillation angle. The highest resolution of the data collected was 1.95 Å, however, the crystals clearly diffracted to higher resolution. Collection of higher resolution data would have required a second pass, to avoid overloaded low order reflections. Due to time constraints 1.95 Å in one pass seemed the best compromise. The data were processed with XDS in the primitive hexagonal space group P6 with cell parameters $a$ = 58.4 Å, $c$ = 32.1 Å (Kabsch, 1993) and later converted to CCP4 format. The reflection intensities were scaled with the CCP4 program SCALA (CCP4, 1994, Evans, 1997) and converted to amplitudes with TRUNCATE (CCP4, 1994).

The derivative data set was collected from a crystal soaked overnight in 5 mM MMA. The crystal was flash–frozen in liquid nitrogen using 15 % glycerol in the motherliquor as cryoprotectant and it was kept at 100 K during data collection. The diffraction data were collected on a Mar345 image plate detector using a copper rotating anode as X–ray source. In order to increase redundancy one hundred frames were collected to 2.37 Å resolution with a one degree oscillation range. The derivative data were processed the same way as for the native using cell parameters $a$ = 58.4 Å, $c$ = 32.1 Å. Native and derivative data processing statistics can be found in Table 5.1.

| ***Data set*** | ***Wavelength (Å)*** | ***Resolution range (Å)*** | ***Total number of reflections*** | ***Number of unique reflections*** | ***Overall $R_{sym}$ (%)*** | ***$R_{sym}$ in the highest resolution shell (1.95–2.06 Å) (%)*** |
|---|---|---|---|---|---|---|
| Native | 1.000 | 50–1.95 | 11984 | 4522 | 5.1 | 17.3 |
| MMA | 1.542 | 30–2.37 | 15224 | 2600 | 4.9 | 10.1 |
| ***Data set*** | ***Overall completeness (%)*** | ***Completeness in the highest resolution shell (1.95–2.06 Å) (%)*** | ***Overall I/σ*** | ***I/σ in the highest resolution shell (1.95–2.06 Å)*** | ***Mosicity (°)*** | |
| Native | 97.6 | 90.9 | 10.5 | 4.2 | 0.8 | |
| MMA | 99.5 | 98.3 | 8.7 | 6.5 | 1 | |

**Table 5.1** Data processing statistics for the native and derivative data of AF–Sm2. $I$, intensity. σ, standard deviation of the intensity. $R_{sym} = (\sum_{hkl} \sum_{i} |(I_{hkl} - \langle I \rangle_h)|) / \sum_{hkl} \sum_{i} |I_{hkl,i}|$ for $i$ observations of a given reflection. <$I$>, mean intensity.

## 5.3 Isomorphous replacement

### 5.3.1 Introduction

As it was shown in chapter 2, the electron density in a crystal can be calculated by the Fourier summation:

$\rho(xyz)=\frac{1}{V}\sum_h\sum_k\sum_l |F(hkl)|\exp[-2\pi i(hx+ky+lz)+i\alpha(hkl)]$ . Since $I(h\ k\ l) \propto |F(h\ k\ l)|^2$, therefore the structure factor amplitude of a given reflection ($h\ k\ l$) can be derived from the measured intensities after applying several correction factors. The last term of the above equation, which is the phase angle of a given reflection, cannot be measured, it has to be derived indirectly. The fastest method to obtain reasonable starting phases is the molecular replacement method (Rossman & Blow, 1962) discussed in chapter 2. However, it requires a structurally similar model to the structure to be solved. Although the increasing number of deposited structures in PDB makes it more and more likely to find a homologous protein structure for molecular replacement, in many cases it is necessary to use MAD or isomorphous replacement to obtain phases.

The isomorphous replacement method was developed by Perutz and co–workers (Green *et al.*, 1954) in order to determine the phase angles of the native protein structure factors. In this method the protein crystal is soaked in a dilute solution of a heavy atom compound (about soaking see Stura & Chen, 1992). There are two major requirements for a successful phase determination by isomorphous replacement: the heavy atom derivative has to bind strong enough to a few sites in the protein molecule and the binding should not cause substantial changes in the crystal structure, in other words the native and derivative crystals should be isomorphous. If the native and derivative crystals are isomorphous, then the structure factors of the heavy atom derivative ($\boldsymbol{F_{PH}}$) are the vectorial sum of the native

protein structure factors ($F_P$) and the heavy atom contribution ($F_H$): $F_{PH}=F_P+F_H$ .

Although $F_H$ cannot be measured directly, it can be calculated if the positions of the heavy atoms are known. The position of the heavy atom sites in the cell can be determined by calculating a Patterson–map with coefficients $|F_{PH}|-|F_P|$ . In its general form the Patterson function is a Fourier summation with zero phase angles and intensities as coefficients:

$P(uvw)=\frac{1}{V}\sum_{hkl}|F(hkl)|^2\cos[2\pi(hu+kv+lw)]$ . The function is calculated at each point $u$, $v$, $w$, of a three dimensional grid which has the same dimensions as the crystal unit cell. The Patterson function has the following properties:

- Peaks in the map represent vectors between atoms in the real unit cell, and every pair of atoms contribute with a peak to the Patterson map.
- The Patterson map is centrosymmetric.
- Screw axes in real space become normal rotation axes in vector space.
- Vectors between atoms related by screw axes or rotation axes appear as peaks, called Harker–peaks localised on certain planes, called Harker–sections.

Once the heavy atom positions in the unit cell have been determined, both the structure factors, $|F_H|$, and the phase, $\alpha_H$, of the heavy atom contribution can be calculated:

$F_H=\sum_{j=1}^{n} q_j f_j \exp[-B_j(\sin^2\Theta)/\lambda^2]\exp[i2\pi(hx_j+ky_j+lz_j)]$ , where $n$ is the number of heavy atom sites, $q_j$ is the occupancy at site $j$, $f_j$ is the atomic scattering factor of the heavy atom at site $j$, $B_j$ is the isotropic temperature factor of the heavy atom at site $j$, $x$, $y$ and $z$ are the coordinates of site $j$ and $h$, $k$ ,$l$ are the Miller indices.

The protein phase angles can now be derived from $|F_P|$, $|F_{PH}|$, $|F_H|$ and $\alpha_H$ in a simple graphical way shown in an Argand diagram in Figure 5.7. This representation is known as the Harker construction. When using only a single derivative there is a phase ambiguity.

which has to be resolved somehow to obtain an interpretable electron density map.

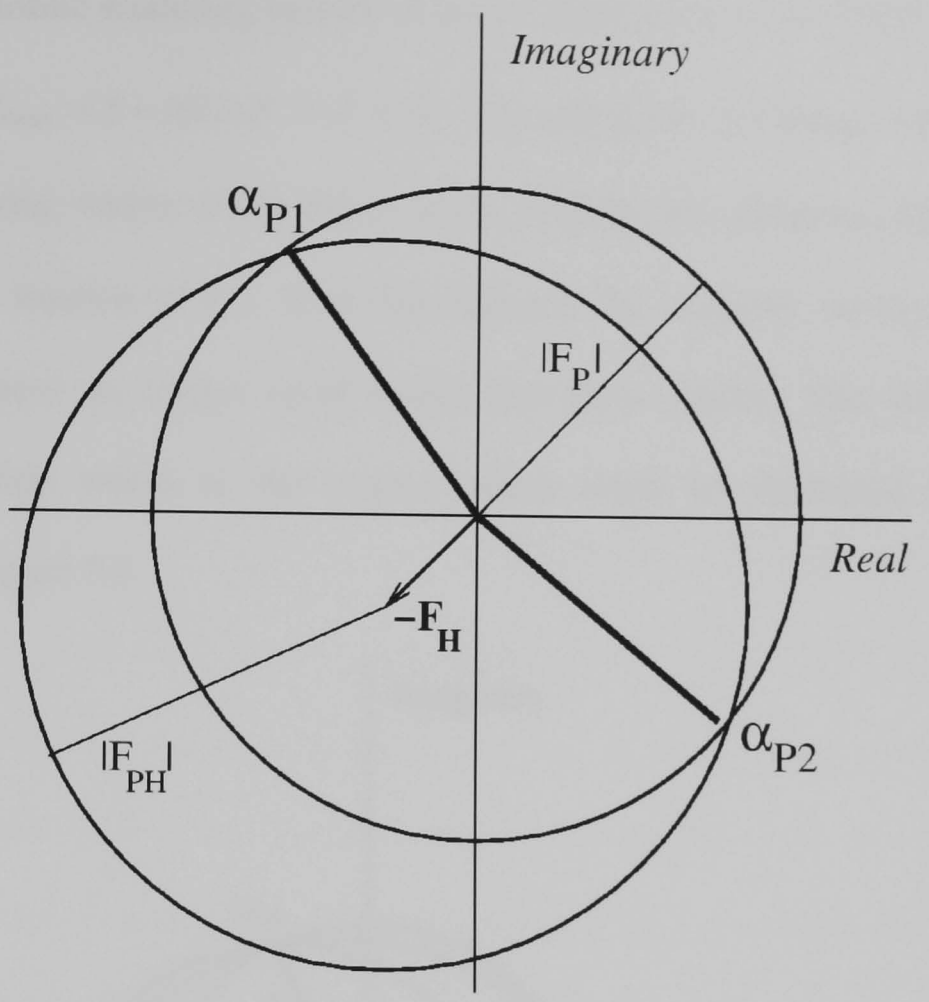



**Figure 5.7** The Harker construction for a single isomorphous derivative. In case of a single derivative there is always an intrinsic phase ambiguity: there are two equivalent intersections of the circles at $\alpha_{P1}$ and $\alpha_{P2}$.

One way of resolving the phase ambiguity of the SIR method is the preparation of further heavy atom derivatives, a method which is called multiple isomorphous replacement (MIR).

There are cases when only a single heavy atom derivative of a protein crystal is available. Even in such a case it is still possible to break the phase ambiguity if anomalous scattering is present. This method is called single isomorphous replacement with anomalous scattering (SIRAS). The inner electrons of the heavy atoms cannot be considered as free electrons therefore they scatter the X–ray beam anomalously, which means that the phase difference between the incident and the scattered beam is not exactly

180 degrees. The strength of anomalous scattering is, however, strongly wavelength dependent. The atomic scattering factors of heavy atoms have to be rather expressed in the following form: $f_{anom.} = f + \Delta f + if'' = f' + if''$. The real term, $\Delta f$ changes only the length of the atomic scattering vector of the heavy atom, whereas the imaginary term, $if''$ causes a counterclockwise rotation of $\boldsymbol{F_H}$. As a consequence, the structure factors $\boldsymbol{F_{PH}}(h\ k\ l)$ and $\boldsymbol{F_{PH}}(-h\ -k\ -l)$ have no longer equal length and phase angles. This difference can be sufficient to decide which is the correct phase angle as illustrated by the Harker construction in Figure 5.8.

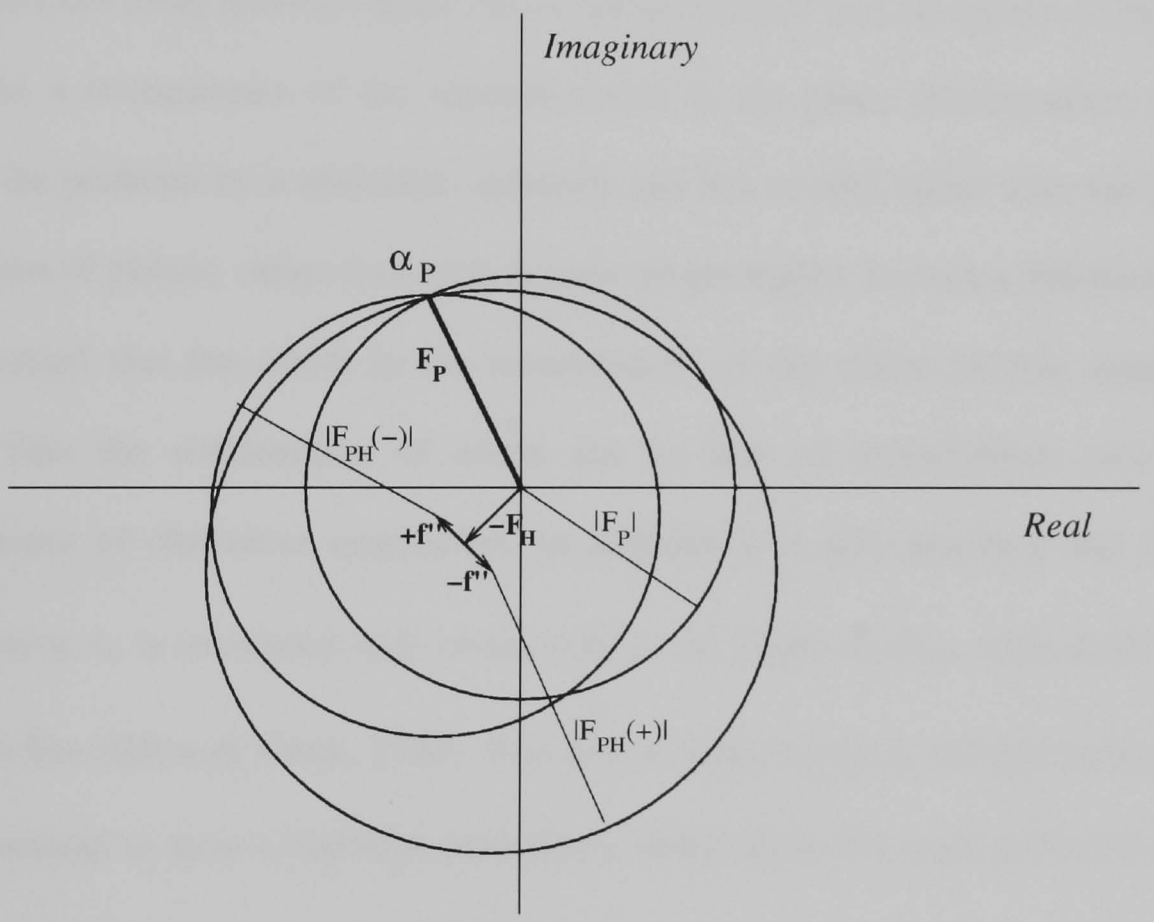



**Figure 5.8** Harker construction showing how anomalous scattering can resolve the phase ambiguity for a single derivative.

The Harker construction in Figure 5.8 shows an ideal case when the three circles intersect in a single point, which is usually not observed when dealing with real data. The measurement of the reflections and the data reduction procedure always implies experimental and computational errors, and there is always a lack of isomorphism between

the native and derivative crystals. In order to calculate the most accurate phase angles and to obtain the best estimates of the phase errors the heavy atom parameters have to be refined (Evans, 1991). The refined parameters are the relative scale and B–factor to put the derivative data on the same scale as the native data, the coordinates, the occupancy and the B–factor of the heavy atoms, and in addition the anomalous occupancy. If the data are good enough, even the anomalous scattering parameters $f'$ and $f''$ can be refined. The target function of the refinement can be minimised by the methods of least–squares, although more and more phasing programs use a maximum likelihood algorithm (Bricogne, 1991), which provides more realistic values for the phasing power and the figures of merit.

As a consequence of the inherent errors in the phase determination one has to address the problem by a statistical approach and has to deal rather with the probability distribution of phases, rather than with discrete phase angles. In such a statistical approach it is assumed that the errors in the measurement of the native protein amplitudes are smaller than the combination of errors due to lack of isomorphism and inaccurate measurement of derivative amplitudes. In addition it is also assumed that the lack of closure error, ε, is associated with errors only in the length of $\boldsymbol{F}_{PH}$, while both $\boldsymbol{F}_H$ and $\boldsymbol{F}_P$ are error–free (Blow & Crick, 1959). The lack of closure $\varepsilon(\alpha)$ is defined as $|\boldsymbol{F}_{PH}|-|\boldsymbol{F}_P+\boldsymbol{F}_H|$ and is assumed to have a Gaussian probability distribution. For each reflection of a given derivative the phase probability is expressed as: $P(\alpha)=P(\epsilon)=N\exp[\frac{\epsilon^2(\alpha)}{2E^2}]$ , where $N$ is a normalisation factor, ε is the lack of closure error and $E^2$ is the mean square of ε. In the case of MIR the phase probability calculated for each derivative can be simply multiplied to obtain a combined phase probability. Hendrickson and Lattman (Hendrickson & Lattman, 1970) proposed an expression for the phase probability, which simplifies the combination of phases calculated from multiple derivatives, including

anomalous data. The phase combination is done simply by adding the coefficients *A*, *B*, *C* and *D* of the following equation for each derivative:

$$P(\alpha)=N\exp(A\cos(\alpha)+B\sin(\alpha)+C\cos(2\alpha)+D\sin(2\alpha)) \quad .$$

The most probable electron density map can be calculated by using phases corresponding to the *maximal* combined probability ($P(\alpha)_{max}$), although this does not necessarily result in the *best* electron density. The reason is, that the combined probability function can have more than a single maximum, which has to be taken into account. Therefore the best electron density map is calculated with $\alpha_{best}$, which is derived from the centroid of the probability distribution rather than from the maximal probability. The best protein structure factor is given by the equation: $F_{P(best)}=m\left|F_{hkl}\right|\exp(i\alpha(best))$ , where *m* is the figure of merit, which is given by $\overline{\cos(\alpha-\alpha(best))}$ , where $\overline{\alpha-\alpha(best)}$ is the estimated error in the phase angle.

## 5.3.2 Isomorphous replacement applied to AF–Sm2

The native and derivative data were scaled together using the CCP4 program FHSCAL and were analysed for isomorphous and anomalous differences by SCALEIT (Table 5.2). An isomorphous difference Patterson map was calculated by FFT with the coefficients $(|F_{PH}| - |F_P|)^2$. The position of a single heavy atom site was located by the real–space Patterson search program RSPS (CCP4, 1994). The Harker section of the difference Patterson map is shown in Figure 5.9. The refinement of the heavy atom parameters and phasing was carried out by SHARP using the heavy atom positions obtained from RSPS. The program implements a maximum likelihood based refinement procedure (Bricogne, 1991; de la Fortelle & Bricogne, 1997). In the first run the temperature factor of the heavy atom was modelled as isotropic, then the second run

refined the anisotropic temperature factors starting with the parameters obtained at the end of the previous run. The phasing statistics are shown in Table 5.2 below. The phase calculation was followed by 99 cycles of solvent flattening and a final cycle of solvent flipping by SOLOMON (Abrahams, 1996). The solvent flattening procedure is embedded in the graphical interface of SHARP.

| Concentration of MMA | 5 mM |
|---|---|
| Number of sites | 1 |
| Number of reflections (centric/acentric) | 269/2358 |
| $R_{iso}$ | 0.245 |
| $R_{Cullis}$ (centric) | 0.502 |
| $R_{Cullis}$ (anomalous) | 0.703 |
| Phasing power (centric/acentric) | 2.56/3.00 |
| Figure of merit (centric/acentric) | 0.633/0.576 |

**Table 5.2** Phasing statistics for AF–Sm2 using a single heavy atom derivative (MMA) with anomalous contribution. The highest resolution used for phase calculation was 2.37 Å. $R_{iso} = \Sigma|F_{PH} - F_P|/\Sigma|F_P|$; $R_{Cullis} = \Sigma||F_{PH} \pm F_P| - |F_H||/\Sigma|F_{PH} \pm F_P|$; Phasing power = r.m.s.($F_H$)/r.m.s.(ε); Figure of merit = $|F_{P(best)}|/|F_P|$. $F_{PH}$, $F_P$ and $F_H$ are the native, derivative and heavy atom structure factor amplitudes respectively, and ε is the lack of closure error.

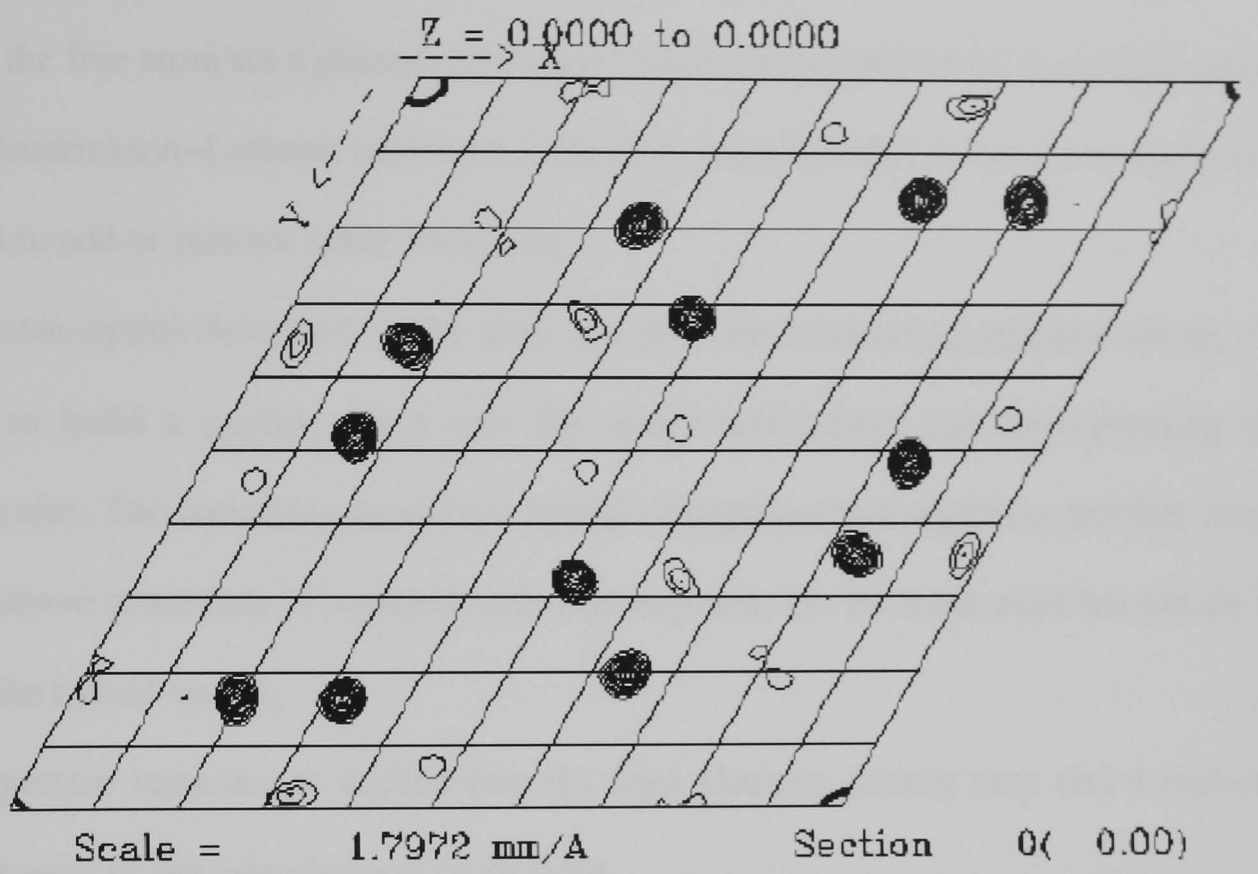



**Figure 5.9** The Harker section at *w*=0 of the isomorphous difference Patterson map contoured from 2σ to 15σ with 1.5σ steps.

## 5.4 Automatic model building and refinement of AF–Sm2

The SIRAS and solvent flattened phases from SHARP and SOLOMON were used to attempt the automatic building and refinement of the structure by the ARP/wARP program suite, version 5.1 (Lamzin & Wilson, 1993; Lamzin & Wilson, 1997; Perrakis *et al.*, 1997; Perrakis *et al.*, 1999). The program suite is a collection of shell scripts and binary programs and it heavily uses CCP4 programs like REFMAC to build the protein structures from scratch. The steps performed in the case of AF–Sm2 were the following:

- In the first step a dedicated shell script helps to set up the variables used in the calculations and outputs a parameter file, which can be easily modified at will.
- A $\sigma_A$–weighted (Read, 1986) map is calculated using the already weighted structure factor amplitudes and the phases from the SOLOMON run, and a so called free atom

model is built into it. The free atom model is essentially a set of water molecules.

- With the free atom set a phased–restrained refinement by REFMAC is carried out using the Hendrickson–Lattman coefficients obtained from SHARP. After every cycle wARP is run to add or remove water molecules.
- After ten cycles described in the previous point an autotracing program (main_trace) tries to build a peptide chain into the model substituting the corresponding water molecules. The remaining model is a hybrid of peptide chain segments and free atoms.
- The above procedure is repeated until convergence, i.e. no more residues can be built into the hybrid model.
- The protein sequence is docked into the final electron density map and a real–space refinement of the side chains is carried out.

The refinement parameters were varied to obtain as complete model as possible, but even in the best case only 64 residues were built into the model. Not surprisingly the missing residues were the three N– and C–terminal residues, residues 24–27 forming loop L2, and residues 52–54 forming loop L4 (see chapter 6).

The structure of AF–Sm2 was refined with the CNS software package (Brünger, 1998) using always the phased maximum likelihood target. The model building was done in Xfit, a program of the Xtalview package (McRee, 1999). Firstly, the residues missing in the ARP/wARP output model were built into the density. The N–terminal residues (1–3) and the L4 loop residues (52–54) could be placed into the density without any problem. The L2 loop residues (24–27) and one more C–terminal residue (75) were built into the map only after a few refinement cycles, initially they were omitted from the model. After 200 cycles of energy minimisation, simulated annealing with a 2500 K starting temperature was run utilising torsion angle dynamics (Rice & Brünger, 1994) to regularise the newly built parts. Successively a group–based B–factor refinement was performed followed by the calculation of $2F_o–F_c$ and $F_o–F_c$ maps. The model was manually rebuilt in Xfit using the real–space fitting capability of the program whenever it was possible. The

geometry of the corrected model was energy minimised followed by individual B–factor refinement, and again electron density maps were calculated. These steps were repeated until manual corrections did improve the model quality (*R*–factor and good fit into the map). The solvent building was done in Xfit with the inspection of each water molecule automatically built by the program. This was done relatively quickly since altogether less then 50 solvent molecules could be located in the asymmetric unit. A larger density feature exactly on the six–fold axis was observed already in the first maps. The globular density is surrounded by the positively charged side chains of K23. These findings and the fact that the crystals were obtained from 2.5 M ammonium sulphate suggested that actually a sulphate ion is located on the six–fold axis and is therefore disordered. In CNS it is not possible to handle atoms on *special positions* covalently linked with other atoms, thus the sulphate was modelled by a single cadmium atom. In addition some waters have been replaced by acetate ions, which better fitted the electron density. The content of the present model and some refinement statistics are shown in Table 5.3.

| **Refinement statistics** | |
|---|---|
| Total number of reflections used | 4521 |
| Working set of reflections | 4295 |
| R–factor (%) | 19.41 |
| Test set of reflections | 226 |
| R–free (%) | 21.13 |
| Total number of protein atoms | 579 |
| Total number of water molecules | 33 |
| Total number of other solvent atoms | 13 |
| **Geometry statistics** | |
| R.m.s.Δ bond distance (Å) | 0.0054 |
| R.m.s.Δ bond angle (°) | 1.32 |
| B–factor R.m.s.Δ ($Å^2$) | |
| Bonded main chain atoms ($Å^2$) | 2.25 |
| Bonded side chain atoms ($Å^2$) | 3.07 |
| Angle main chain atoms ($Å^2$) | 3.36 |
| Angle side chain atoms ($Å^2$) | 4.73 |
| Average B factor ($Å^2$) | |
| Main chain atoms ($Å^2$) | 19.61 |
| Side chain atoms ($Å^2$) | 21.48 |
| All protein atoms ($Å^2$) | 22.07 |
| Water molecules ($Å^2$) | 37.86 |
| Other solvent atoms ($Å^2$) | 46.78 |

**Table 5.3** Refinement data of the AF–Sm2 model.

## 5.5 Validation of the model

The electron density map calculated from the final model was of good quality without unexplained positive or negative difference density features. The quality of the model was analysed by the structure validation programs PROCHECK (Laskowski *et al.*, 1993) and WHATIF (Vriend, 1990). WHATIF suggested only minor corrections to the model concerning side chain torsion angle conventions. The Ramachandran plot calculated

by PROCHECK is shown on Figure 5.10.

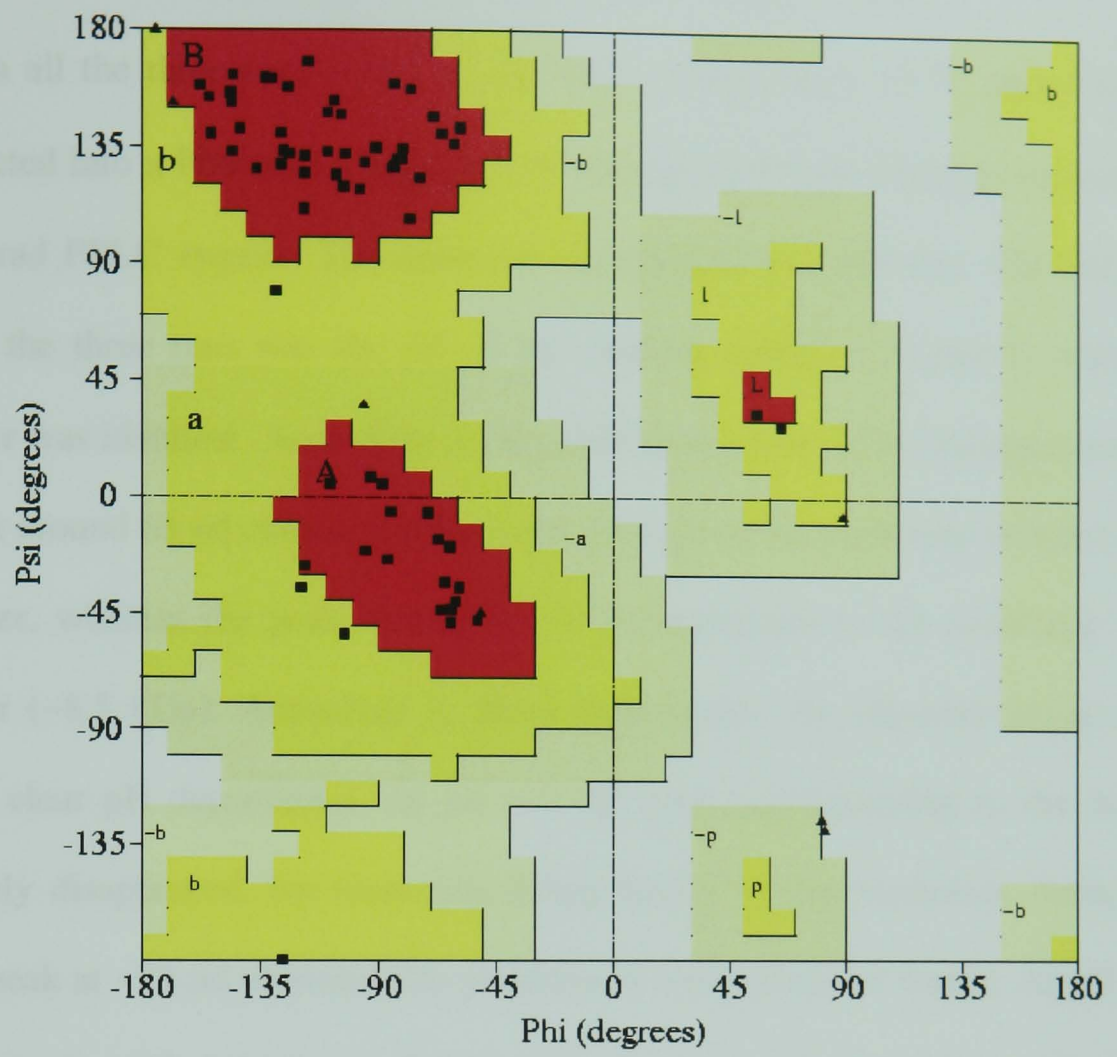



**Figure 5.10** Ramachandran plot for the refined model of AF–Sm2. There are no residues in the disallowed region of the plot (white). Most of the residues have phi–psi values in the most favoured region (red), the rest are in the allowed regions (bright yellow). Glycine residues are represented by black triangles, the phi–psi value of other residues are shown as black squares.

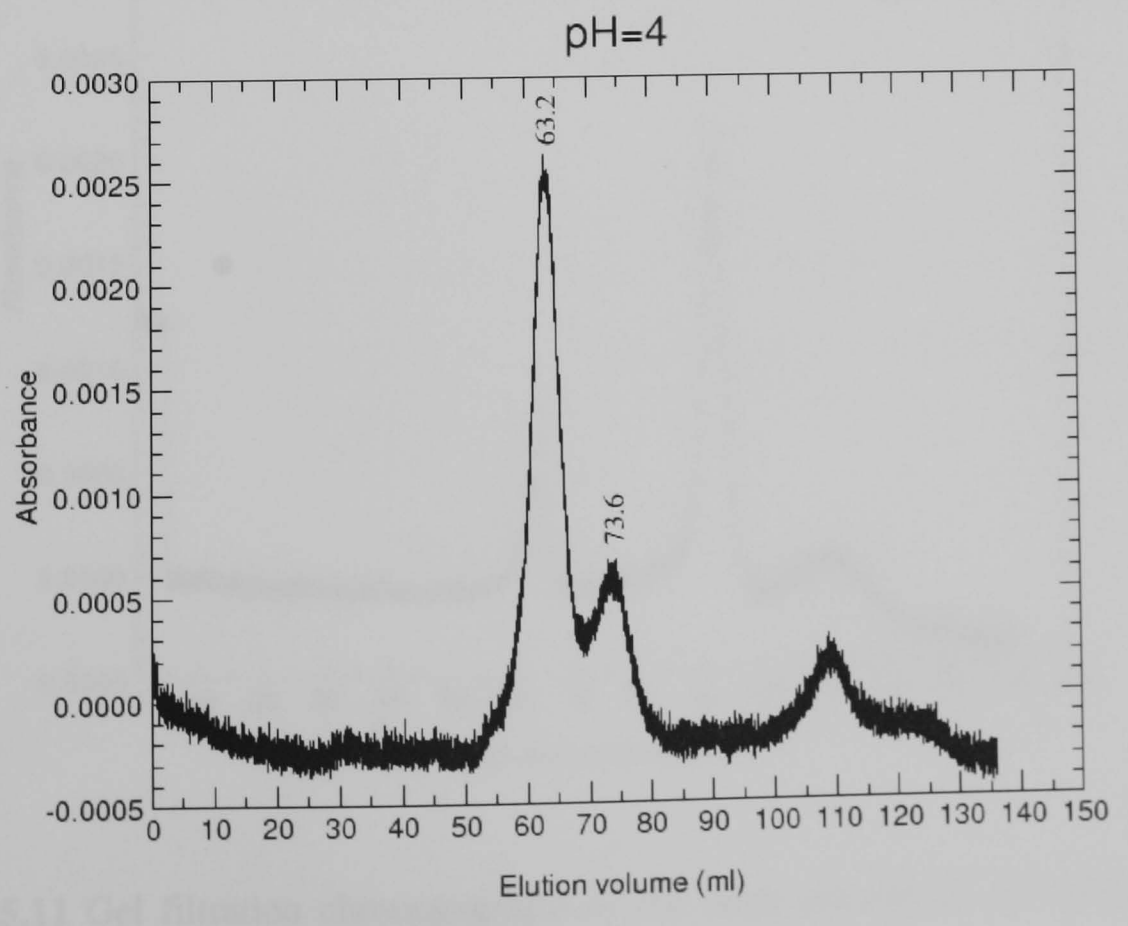
## 5.6 Oligomerisation of AF–Sm2 in solution

As a consequence of the hexagonal symmetry in the crystal lattice the protein molecules are arranged in a hexameric ring. The interactions between the monomers seen in the hexameric model, which will be discussed in the next chapter, strongly suggested that the packing in the crystal is not an artefact, since AF–Sm2 forms hexamers or multimer aggregates also in solution. In order to see the hexamer formation in solution gel

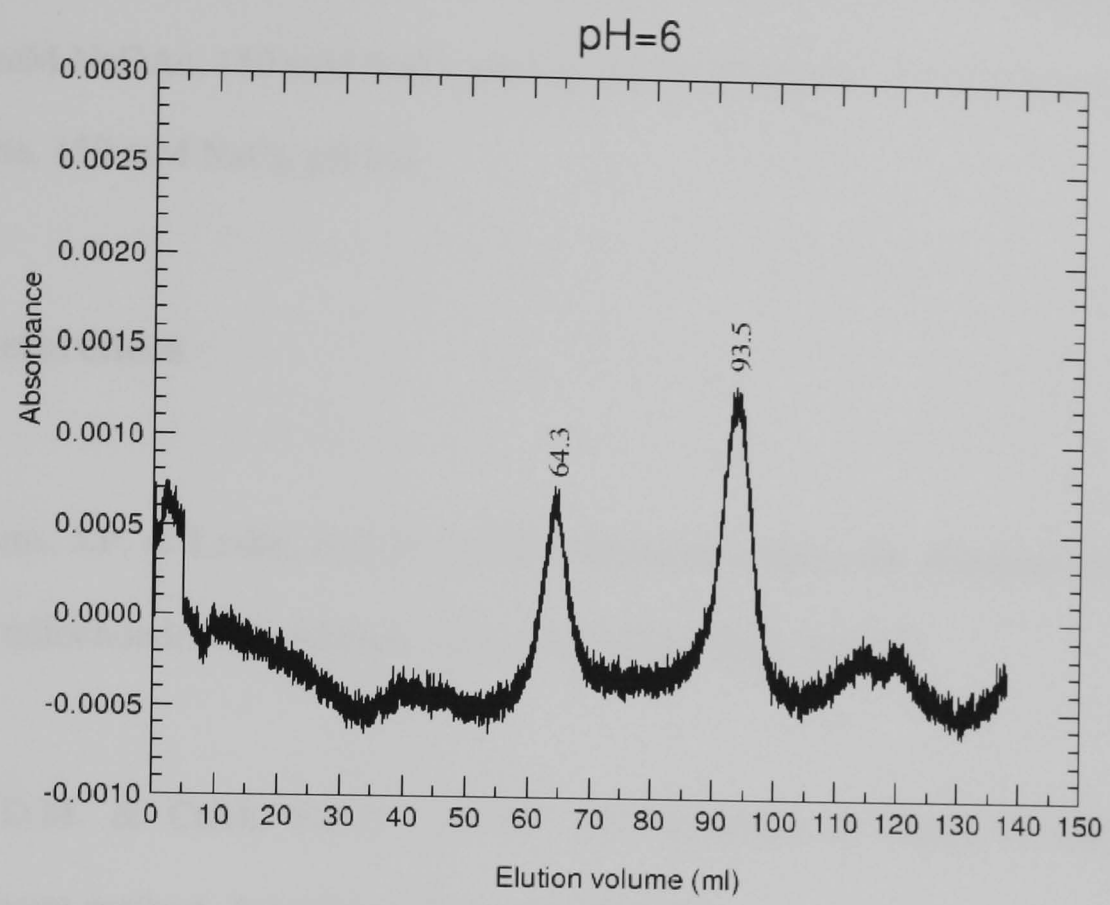
filtration experiments were carried out at three different pH values.

In all the three runs 10 µl of AF–Sm2 solution with 11.15 mg/ml concentration was injected into a Pharmacia Superdex 75 HiLoad 16/60 gel filtration column connected to a Biorad FPLC system. The detection wavelength was 280 nm. The only difference between the three runs was the pH of the running buffer: 4, 6 and 8, otherwise every parameter was identical. According to the calibration curve of the column (not shown) the first peak around 63 ml elution volume corresponded to the molecular weight (~50 kDa) of a hexamer, whereas the peak around 93 ml corresponded to the molecular weight of a monomer (~8.5 kDa). According to these experiments the oligomerisation of AF–Sm2 shows a clear pH dependence. At pH 4.0 the peak corresponding to the monomer has completely disappeared, the molecules being mostly in the hexameric form, although a smaller peak at ~73 ml indicated the presence of some dimer or trimer. At pH 6.0 and pH 8.0 hexameric and monomeric forms were present with increasing monomer content towards higher pH (Figure 5.11).

**A**



**B**




**C**

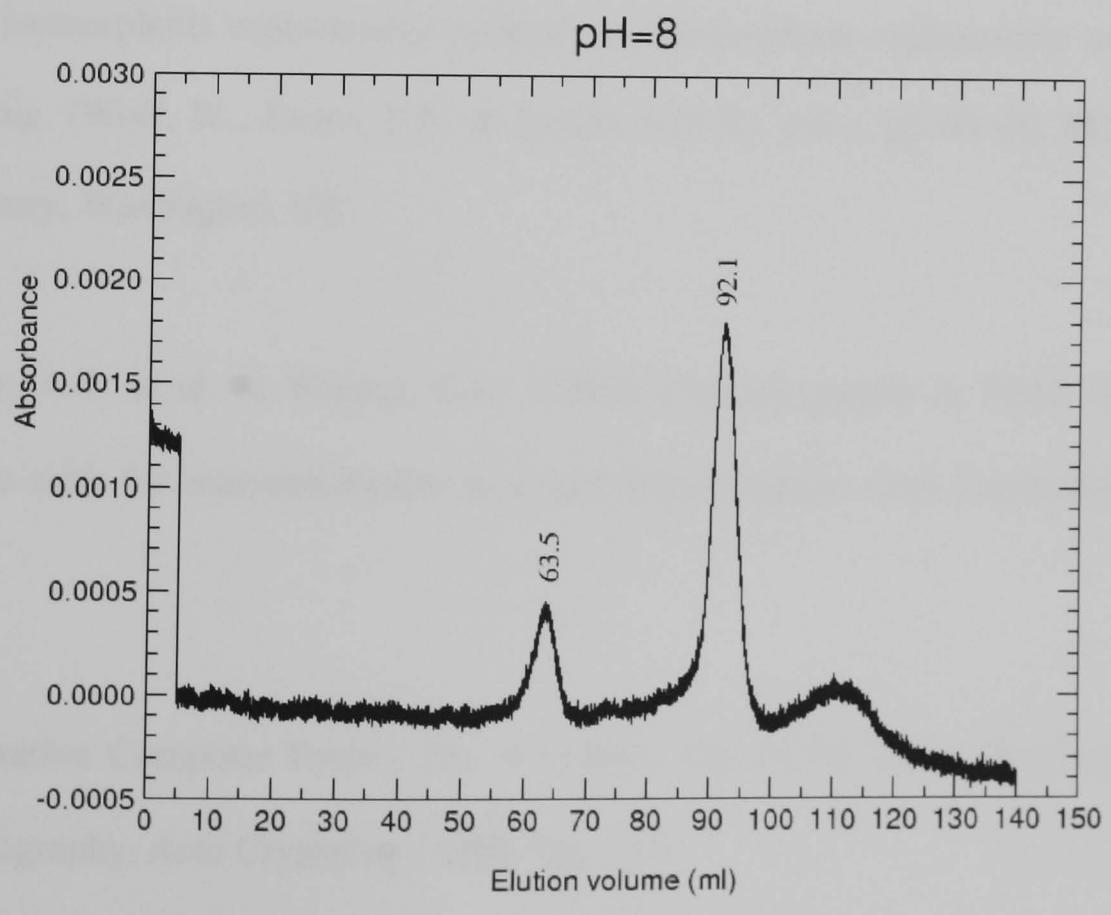



**Figure 5.11** Gel filtration chromatograms of the same amount of AF–Sm2 run at three

different pH. Three different peaks were detected at ~93, ~73 and ~63 ml corresponding to hexamerix, tri– or dimeric and monomeric form respectively. The running buffers were: **A)** 20 mM NaOAc, 150 mM NaCl, pH 4.0; **B)** 20 mM MES, 150 mM NaCl, pH 6.0; **C)** 20 mM Tris, 150 mM NaCl, pH 8.0.

## 5.7 References

Abrahams, J.P. & Leslie, A.G.W. (1996). Methods used in the structure determination of bovine mitochondrial $F_1$ ATPase. *Acta Crystallogr.* **D52**, 30–42.

Blow, D.M. & Crick, F.H.C. (1959). The treatment of errors in the isomorphous replacement method. *Acta Crystallogr.* **12**, 794–802.

Bricogne, G. (1991). A maximum likelihood theory of heavy atom parameter refinement in the isomorphous replacement method. In *Isomorphous replacement and anomalous scattering.* (Wolf, W., Evans, P.R. & Leslie, A.G.W., eds.), pp 60–68, SERC Daresbury Laboratory, Warrington, UK

Brünger, A.T. et al. & Warren, G.L. (1998). Crystallography & NMR System: a new software suite for macromolecular structure determination. *Acta Crystallogr.* **D54**, 905–921.

Collaborative Computer Project No. 4 (1994). The CCP4 Suite: Programs for Protein Crystallography. *Acta Crystallogr.* **D50**, 760–763

De la Fortelle, E. & Bricogne, G. (1997). Maximum–likelihood heavy atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. *Methods Enzymol.* **276**, 472–494.

Evans, P.R. (1991). Refienement of heavy–atom parameters and isomorphous phasing. In *Isomorphous replacement and anomalous scattering.* (Wolf, W., Evans, P.R. & Leslie, A.G.W., eds.), pp 49–59, SERC Daresbury Laboratory, Warrington, UK

Evans, P.R. (1997). Scala. *Joint CCP4 and ESF–EACBM Newsletter* **33**, 22–24.

Green, D.W., Ingram, V.M. & Perutz, M.F. (1954). *Proc. R. Soc. London, Ser. A* **225**, 287

Hendrickson, W.A. & Lattman, E.E. (1970). Representation of phase probability distributions for simplified combination of independent phase information. *Acta Crystallogr.* **B26**, 136–143.

Kabsch, W. (1993). Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J. Appl. Cryst.* **26**, 795–800.

Klenk, H.–P. *et al.* (1997). The complete genome sequence of the hyperthermophilic, sulphate–reducing archaeon *Archaeoglobus fulgidus*. *Nature*, **390**, 364–370.

Lamzin, V.S. & Wilson, K.S. (1993). Automated refienement of protein models. *Acta Cryst.* **D49**, 129–149.

Lamzin, V.S. & Wilson, K.S. (1997). Automated refienement for protein crystallography. *Methods Enzymol.* **277**, 269–305.

Laskowski, R.A., MacArthur, M.W., Moss, D.S. & Thornton, J.M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26**, 283–291.

Lowry, O.H. et al. (1951). *J. Biol. Chem.* **193**, 265.

McRee, D.E. (1999). XtalView/Xfit – A versatile program for manipulating atomic coordinates and electron density. *J. Struct. Biol.* **125**, 156–165.

Perrakis, A. Sixma, K., Wilson, K.S. & Lamzin, V.S. (1997). wARP: improvement and extension of crystallographic phases by weighted averaging of multiple refined dummy atom models. *Acta Cryst.* **D53**, 448–455.

Perrakis, A., Morris, R.J. & Lamzin, V.S. (1999). Automated protein model building combined with iterative structure refienement. *Nature Struct. Biol.* **6**, 458–463.

Read, R.J. (1986). Improved Fourier coefficient s for maps using phases from partial structures with errors. *Acta Crystallogr.* **A42**, 140–149.

Rice, L.M. & Brünger, A.T. (1994). Torsion angle dynamics: reduced variable conformational sampling enhances crystallographic structure refinement. PROTEINS: *Structure, Function and Genetics* ***19***, 277–290.

Rossmann, M.G. & Blow, D.M. (1962). The detection of sub–units within the crystallographic asymmetric unit. *Acta Crystallogr.* **15**, 24–31.

Studier, F.W., Rosenberg, A.H., Dunn, J.J. & Dubendorff, J.A. (1990). Use of T7 RNA polimerase ti direct expression of cloned genes. *Methods Enzymol.* **185**, 60–89.

Stura, E.A. & Chen, P. (1992). Soaking of crystals. In *Crystallization of nucleic acids and proteins. A practical approach.* Ducroix, A. & Giegé, R., Eds., Oxford University Press, New York, pp. 311–317.

Vriend, G. (1990). WHAT IF: a molecular modelling and drug design program. *J. Mol. Graph.* **8**, 52–56.

Zeelen, J.Ph., Hiltunen, J.K., Ceska, T.A. & Wierenga, R.K. (1994). Crystallisation experiments with 2–enoyl–CoA hydratase, using an automated 'fast–screening' crystallisation protocol. *Acta Crystallogr.* **D50**, 443–447.

# Chapter 6

# Structure analysis of AF–Sm2 from *Archaeoglobus fulgidus*

## 6.1 Introduction

This chapter describes the molecular structure of an Sm–like protein cloned from *Archaeoglobus fulgidus*. The structure has been solved by SIRAS. At present the function of this archaebacterial Sm–like protein is unknown, although there is an ongoing effort to identify its targets in the archaeon *A. fulgidus*. The crystal structure of two complexes of human spliceosomal Sm proteins have recently been determined (Kambach *et al.*, 1999) giving us the opportunity to make a structural comparison of the constituent monomers, as well as the interactions within the complexes. The structure of one of the two Sm–like proteins present in *Archaeoglobus fulgidus*, named AF–Sm2 will be discussed here and compared with the structure of the human Sm proteins (Kambach *et al.*, 1999).

## 6.2 Quality of the model

The current model contains 75 residues of the 77 total. The two C–terminal glutamate residues (E76 and E77) are not included in the current model as they are completely disordered and are not visible in the electron density map even at low contour levels (<1$\sigma$), whereas the N–terminal residues have well defined electron density (Figure 6.2). The geometry of the refined structure, as has been summarised in Table 5.3 of the previous chapter, can be considered good. The residues comprising the N–terminal $\alpha$ –helix and the five stranded $\beta$–sheet have lower (~20 Å$^2$) main chain temperature factors and form the more rigid core of the molecule. As expected, the turns and loops connecting

the antiparallel β–strands are more flexible and their main chain B–factors refine to higher values in the range 40–50 $Å^2$ (Figure 6.1A and B).

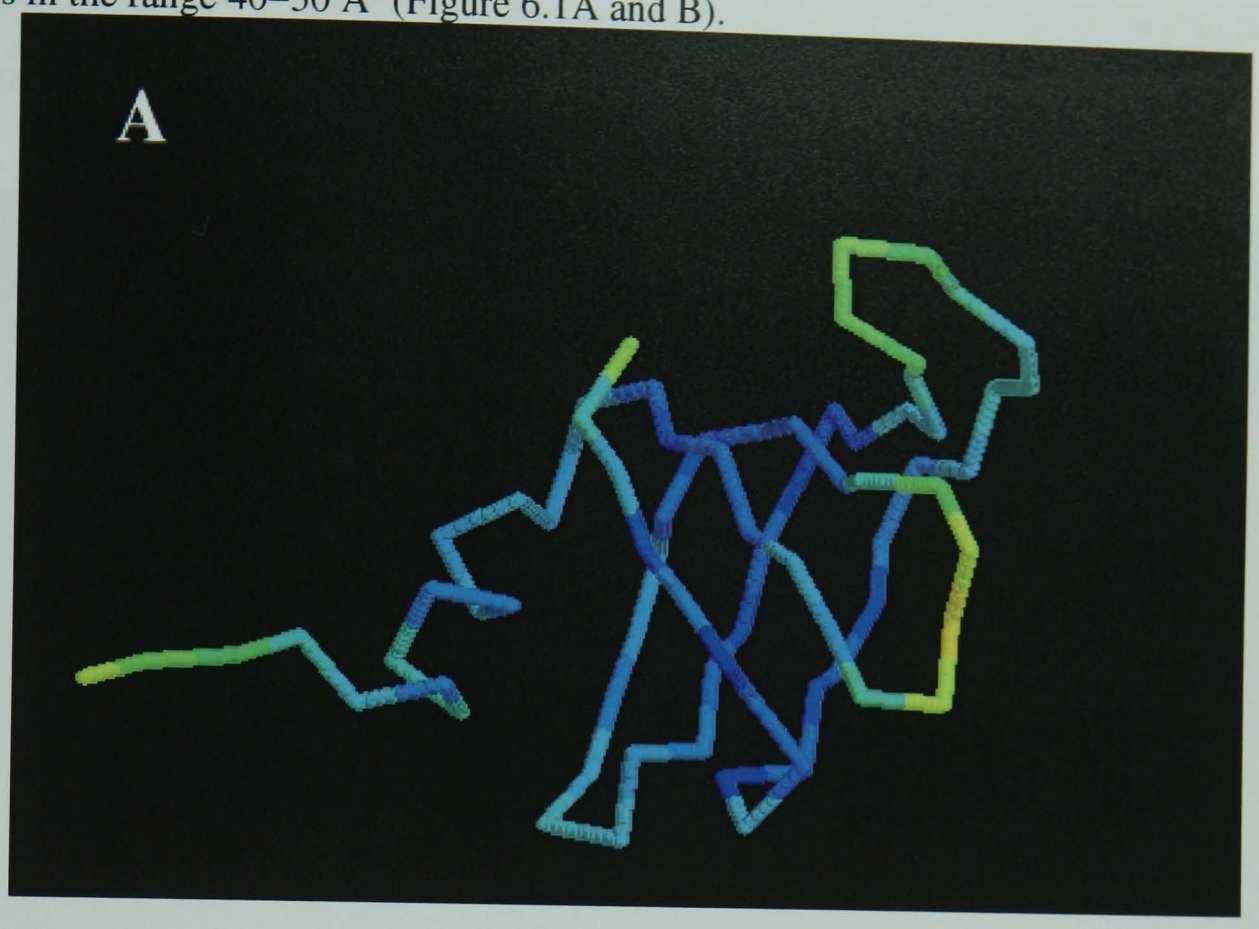



**Figure 6.1 A)** The $C_\alpha$ trace of AF–Sm2 coloured according to the main chain B–factors. Low B–factors are indicated by a blue colour. The five β strands and the N–terminal helix form a rigid core of the molecule. The N–terminal extension and two loops (L2 and L4,

see Figure 6.3) have significantly higher thermal fluctuations. This figure was prepared by RASMOL version 2.7.1 included in the CCP4 suite. **B)** The main chain B–factors plotted as a function of the residue numbers.

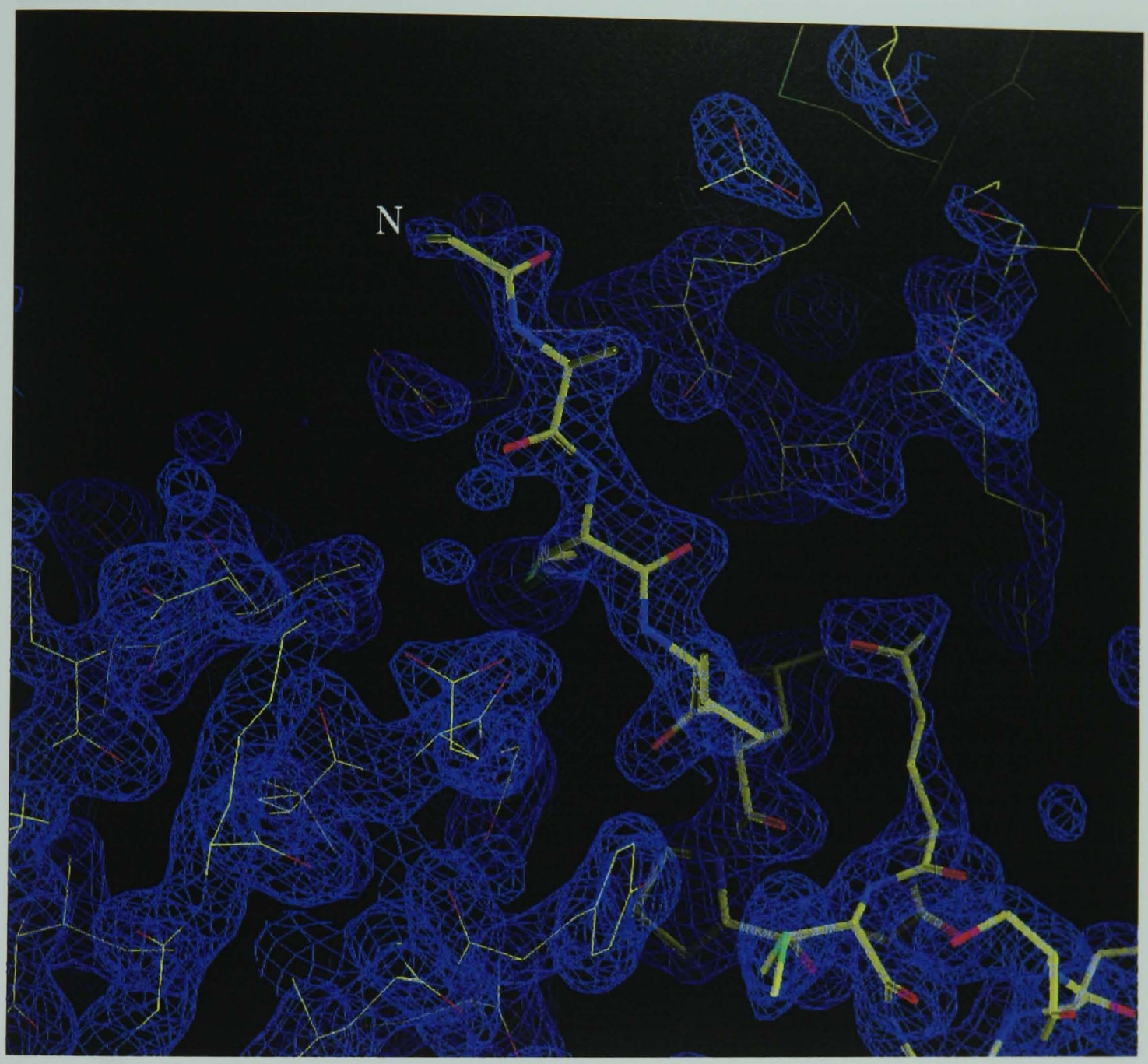



**Figure 6.2** A $2F_o–F_c$ electron density map of AF–Sm2 showing the N–terminus of the molecule. Symmetry related molecules are drawn with thin lines.

In addition to the peptide chain the model contains 33 water molecules, 3 acetate ions and a sulphate ion sitting on the crystallographic six–fold axis. In order to properly model the disordered sulphate ion one could place it into the model with its three–fold axis exactly aligned with the crystallographic six–fold. As a result the sulphur and an oxygen atom of the sulphate would be placed exactly on the six–fold axis, therefore one should

assign an occupancy of 1/6 for these atoms. The remaining oxygen atoms of the sulphate could be arranged around the six–fold axis according to six–fold symmetry assigning half occupancy to each atoms. Since atoms at special positions linked covalently to other atoms cannot be refined in CNS therefore the sulphate was modelled by a spherically symmetric cadmium atom. In total the model contains five atoms on special positions.

## 6.3 The overall structure of AF–Sm2: the Sm fold

The alignment of the available Sm protein sequences (in 1995) identified two highly conserved regions, termed *Sm1* and *Sm2* motifs (Hermann *et al.*, 1995; Séraphin, 1995). Sm proteins form the core protein domain of eukaryotic spliceosomal snRNPs U1, U2, U4 and U5 with the exception of U6. However, an Sm–like protein was found in yeast associated with U6 snRNA possessing high sequence similarity, especially within the Sm motifs, to the already known canonical Sm protein sequences (Cooper *et al.*, 1995, Séraphin, 1995). Sensitive database searches revealed additional Sm–like proteins also in the archaeal domain (Salgado–Garrido *et al.*, 1999). The sequence alignment of the presently known Sm and Sm–like proteins includes more than 80 sequences and is shown in Appendix B.

The crystal structures of four human Sm proteins forming two dimeric complexes (B and $D_3$, $D_1$ and $D_2$) were published last year revealing a new, distinct fold as a hallmark of the Sm and Sm–like proteins (Kambach *et al.*, 1999). The Sm fold consists of an N–terminal α–helix followed by a strongly bent five–stranded antiparallel β–sheet resulting in a barrel–like shape. The *Sm1* motif is formed by strands β1, β2 and β3, while the *Sm2* motif is formed by strands β4 and β5. The two motifs are linked by the L4 loop, which is relatively short in AF–Sm2 (Figure 6.3).

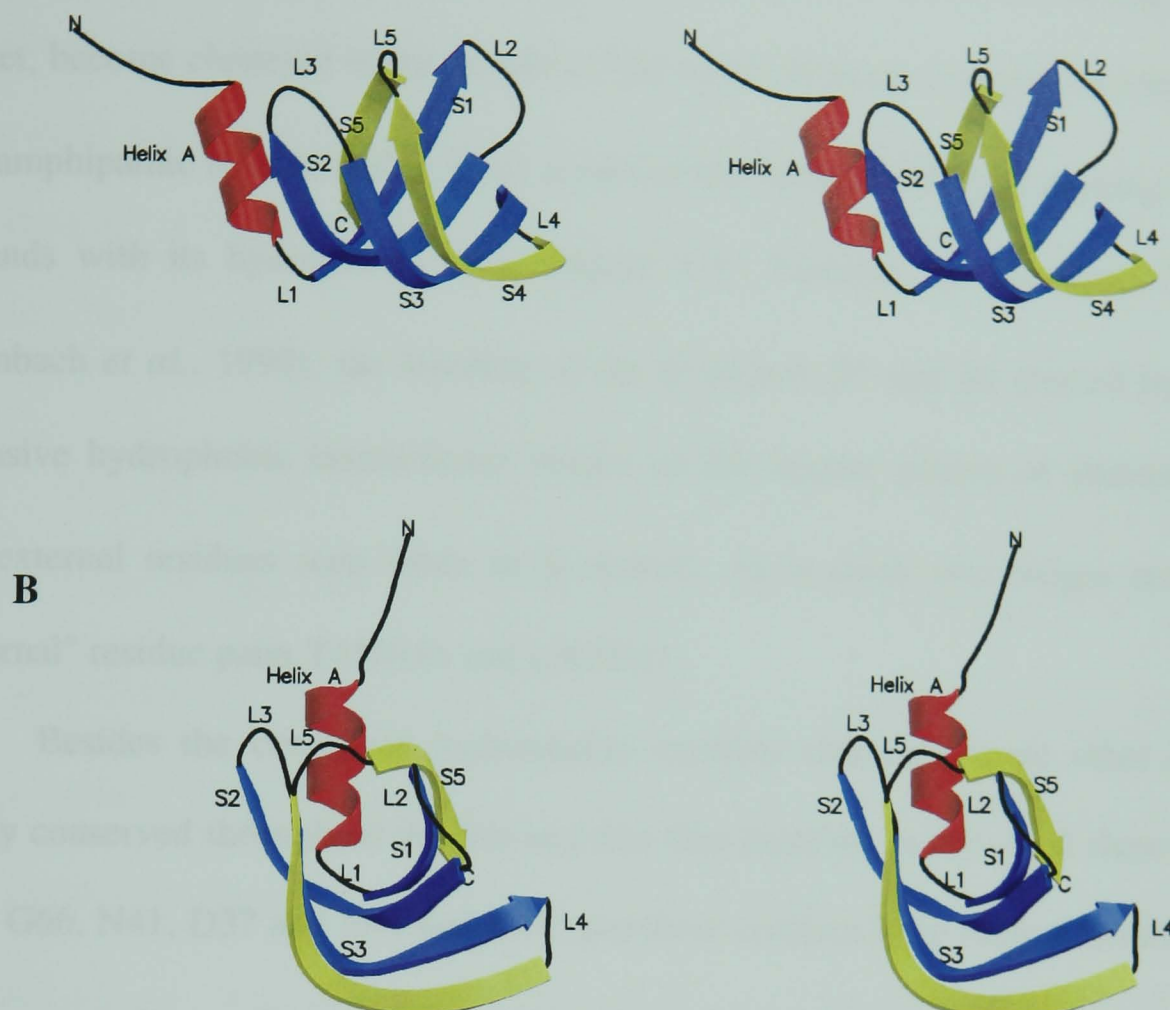



**Figure 6.3** Stereo pictures of the ribbon representation of the AF–Sm2 structure in two orientations. The label of β–strands starts with letter S, the loops have labels starting with letter L. The N–terminal α–helix is coloured red, the β–strands forming the *Sm1* and *Sm2* motifs are blue and yellow respectively. The figure was prepared using the programs MOLSCRIPT version 2.1.2 (Kraulis, 1991) and RASTER3D version 2.4j (Merritt & Bacon, 1997).

In Sm and Sm–like proteins there are a dozen residues which are conserved in more than 2/3 of all cases. Most of these highly conserved residues have hydrophobic side chains. Strands β2, β3 and β4 are strongly bent in all known Sm and Sm–like structures. The severe bending of strand β2 is facilitated by a conserved glycine (G31) providing

greater conformational variation of the main chain. A consequence of such a strong bending is that the hydrophobic side chains, which point towards the internal side of the β –sheet, become clustered in the middle of the barrel forming a compact hydrophobic core. The amphipathic N–terminal α–helix is part of the hydrophobic core packing against the β –strands with its hydrophobic face (Figure 6.9). Similarly to the human Sm proteins (Kambach *et al.*, 1999), the bending of the β–strands β3 and β4 (forced by β2 and the extensive hydrophobic interactions) breaks up the regular pattern of alternating internal and external residues seen often in β–strands. As a result two bulges are formed by "external" residue pairs T45/N46 and G60/E61.

Besides the conserved hydrophobic residues there are some other amino acids highly conserved throughout the Sm and Sm–like proteins. In AF–Sm2 these residues are G31, G66, N41, D37 and E49 and R65, and these residues have clear structural role. G31, as discussed above, allows the bending of the β2 strand. Residues D37, N41, G66 and R65 participate in an inter–subunit hydrogen bonding network (see Figure 6.8 for G66 and R65).


**Figure 6.4** The sequence of AF–Sm2 showing the corresponding secondary structure elements (Figure 6.3) and conserved residues as well. Helix A is coloured red, the *Sm1* motif is blue (β–strands β1, β2 and β3) and the *Sm2* motif is yellow (β–strands β4 and β5). The hydrophobic residues conserved in at least 2/3 in the known Sm and Sm–like

proteins are shown in blue boxes. The following non–hydrophobic residues are almost completely invariant in the Sm and Sm–like proteins: G31, G66 (orange), D37, E49 (magenta), N41 (green) and R65 (red). The C–terminal residues in grey–shaded box are disordered and missing from the model. The two N–terminal residues which are cloning artefacts are indicated in a box of cyan colour. The figure was created with ALSCRIPT version 2.0.5 (Barton, 1993).

The superposition of the AF–Sm2 structure and the four human Sm protein structures shows high structural similarity between these representatives of eukaryotic Sm and archaebacterial Sm–related proteins. As shown in Figure 6.5 the β–strands and even the α–helices (with the exception of $D_2$) superimpose quite well. Indeed, the Sm motifs are the structurally most invariant parts of these proteins, whereas the L4 loop and the C–termini can vary considerably. In AF–Sm2 the $C_\alpha$ positions of two residues in strand β2 close to the L2 loop deviate most from the $C_\alpha$ positions of the same residues in the human Sm structures (Figure 6.5). These two residues break the β–strand and form a short coil.

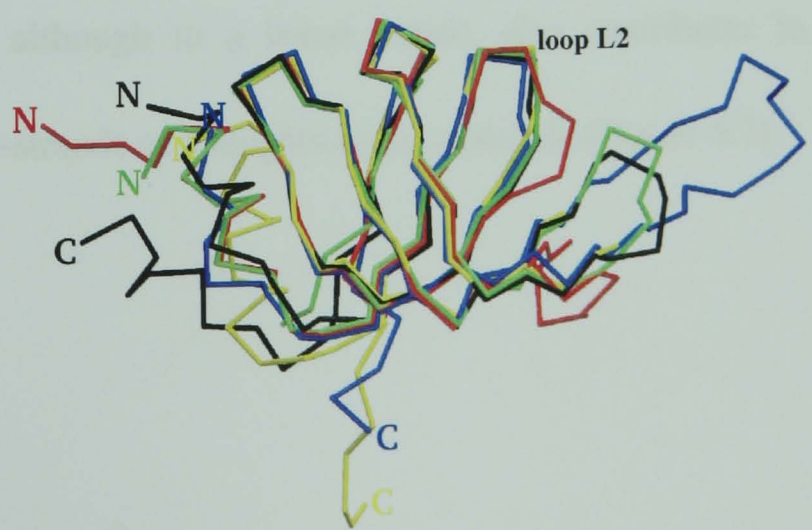



**Figure 6.5** Superposition of the four human Sm proteins B (blue), $D_1$ (black), $D_2$ (yellow)

and $D_3$ (green) with AF–Sm2 (red). The $C_\alpha$ atoms in strands β2, β3, β4 and β5 were used in the superposition showing an r.m.s.Δ less than 0.9 Å. The main chain of AF–Sm2 deviates from the human Sm proteins just after loop L2, forming a β–bulge.

## 6.4 The oligomerisation of AF–Sm2

The asymmetric unit of the AF–Sm2 crystals contain a single protein molecule. The molecules are ordered around a crystallographic six–fold axis forming a hexamer as a consequence of the crystal symmetry. However, the inspection of the molecular organisation of AF–Sm2 monomers in the lattice indicates that the hexamer formation is probably not an artefact of the crystal packing but actually that the crystal is built from pre–existing hexamers in solution (Figure 6.6). Indeed, a pH dependent hexamer formation was demonstrated by gel filtration experiments indicating almost exclusive presence of the hexameric form at the pH of crystallisation (Figure 5.11).

The hexameric ring formed by six AF–Sm2 molecules is a homo–hexamer therefore there is a unique interaction interface between the building blocks. The major interaction interface between neighbouring monomers involves β–strands β4 and β5. The N–terminal α–helix, although to a lesser extent, also contributes to the interface by interacting with the β–strands of a neighbouring molecule (Figure 6.7).

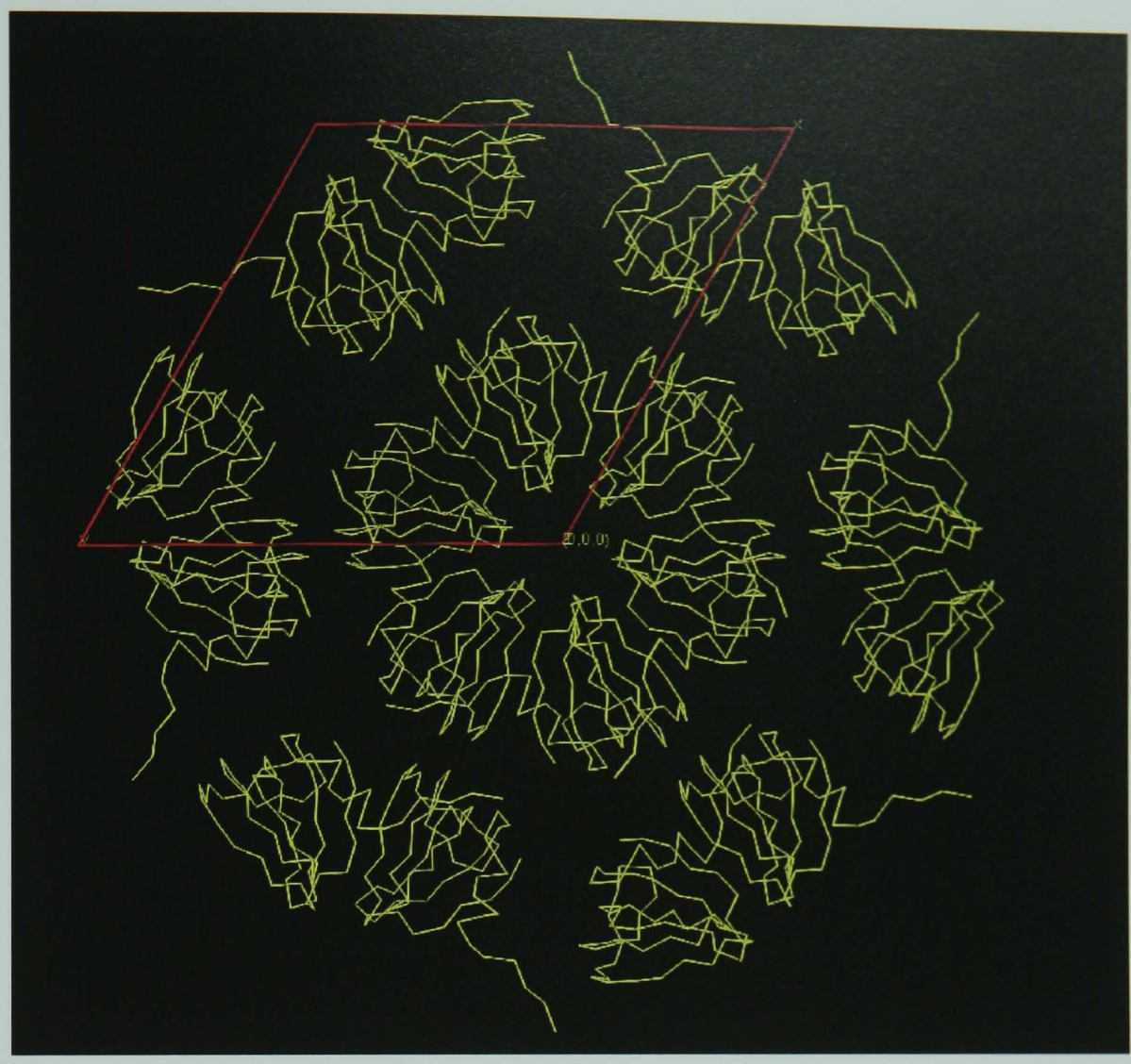


**Figure 6.6** The packing of AF–Sm2 molecules in the hexagonal crystal lattice. The *z* axis of the unit cell (red) is perpendicular to the sheet. The figure shows the interaction of the β –strands of neighbouring molecules forming a continuous circular β–sheet. The formation of a similar circular β–sheet was found in TRAP (Antson *et al.*, 1999).

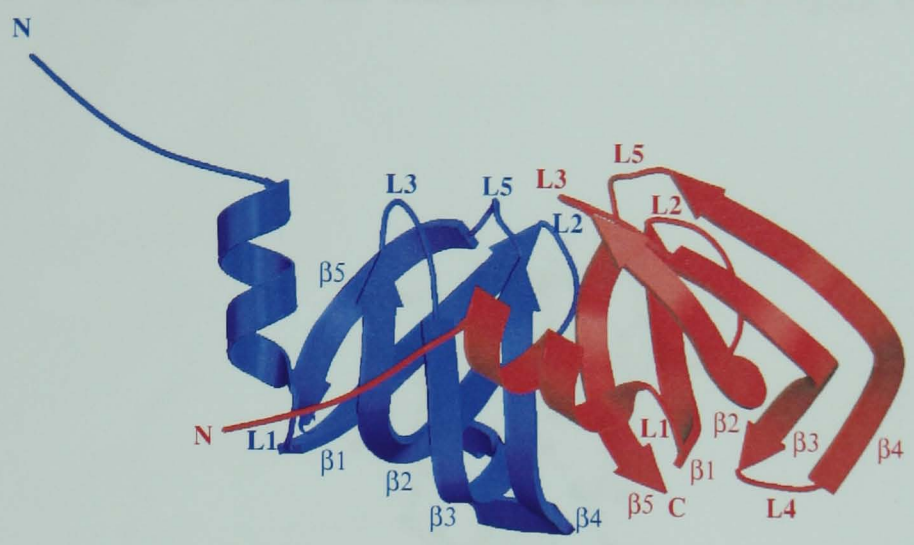



**Figure 6.7** The interaction between two neighbouring monomers in the crystal lattice. The major interaction interfaces are the β4 strand of one monomer (blue), and the β5 strand of the other (red). The N–terminal α–helix (red) packing against strands β4, β3 and β2 (blue) also contributes to the interactions resulting in oligomerisation.

The interaction between strands β4 and β5 is similar to that observed in regular β–sheets, which involves hydrogen bonding between N–H and C=O groups of the main chain and (mostly) hydrophobic interactions between the side chains exposed on one side of the β –sheet (Figure 6.8). As a result of the six–fold symmetry, the interaction maintained by the β–strands β4 and β5 creates a continuous β–ring throughout the hexamer (Figure 6.6). Hydrophobic contacts between the N–terminal helix of one monomer with the β–strands of its neighbour increases the stability of the hexamer (Figure 6.9). In addition there is a single hydrogen bond between the main chain atoms of E34 and V4 (yellow, Figure 6.9).

In summary, the interactions found between monomers of the AF–Sm2 hexamer and the dimers of human Sm proteins are essentially the same, and they result in a very similar spatial arrangement of the interacting monomers (Figure 6.7; Kambach *et al.*, 1999).

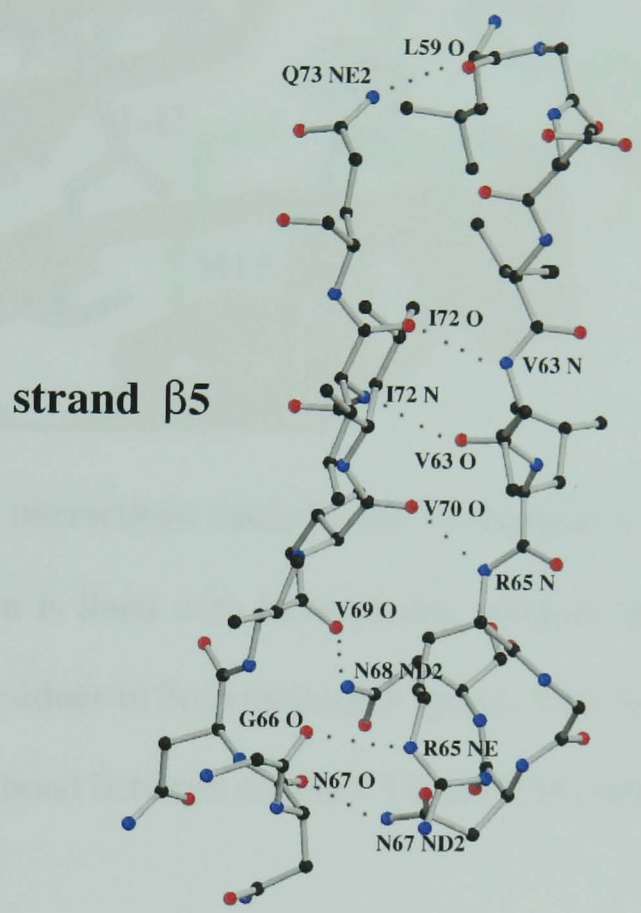



**Figure 6.8** The interactions between β–strands β4 and β5 in two neighbouring molecules. The middle of the strands contain hydrophobic residues making classical (β–sheet) hydrogen bonding interactions between main chain polar groups N–H and C=O (V63, V70, V72). At both ends of the strands, where the main chain distance is increased, the hydrogen bonds are mediated by the side chains of polar residues (N67, R65, Q73).

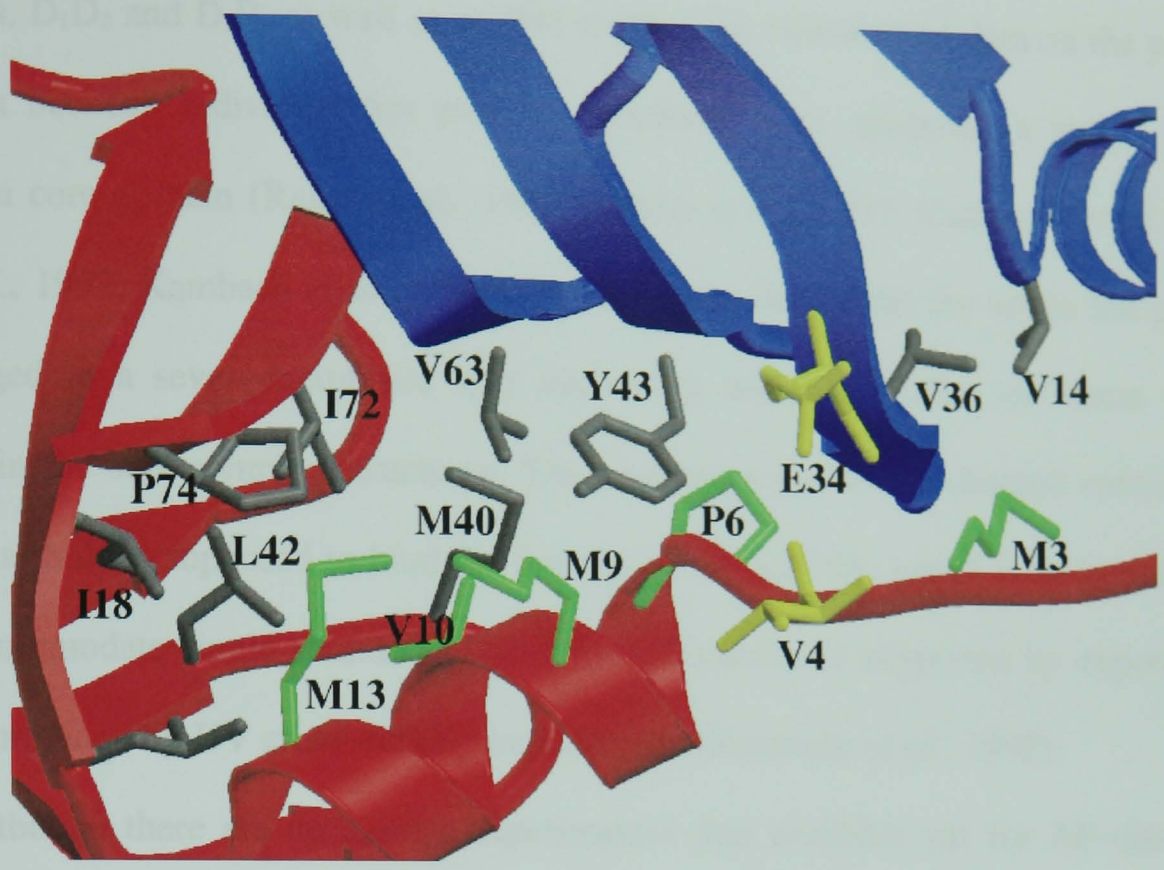



**Figure 6.9** The interactions made by the N–terminal α–helix. The upper side of the helix and its extension is lined with hydrophobic residues (green) packed against hydrophobic side chains of residues in both molecules (grey). This interaction is strengthened by a main chain hydrogen bond between residues V4 and E34 (yellow).

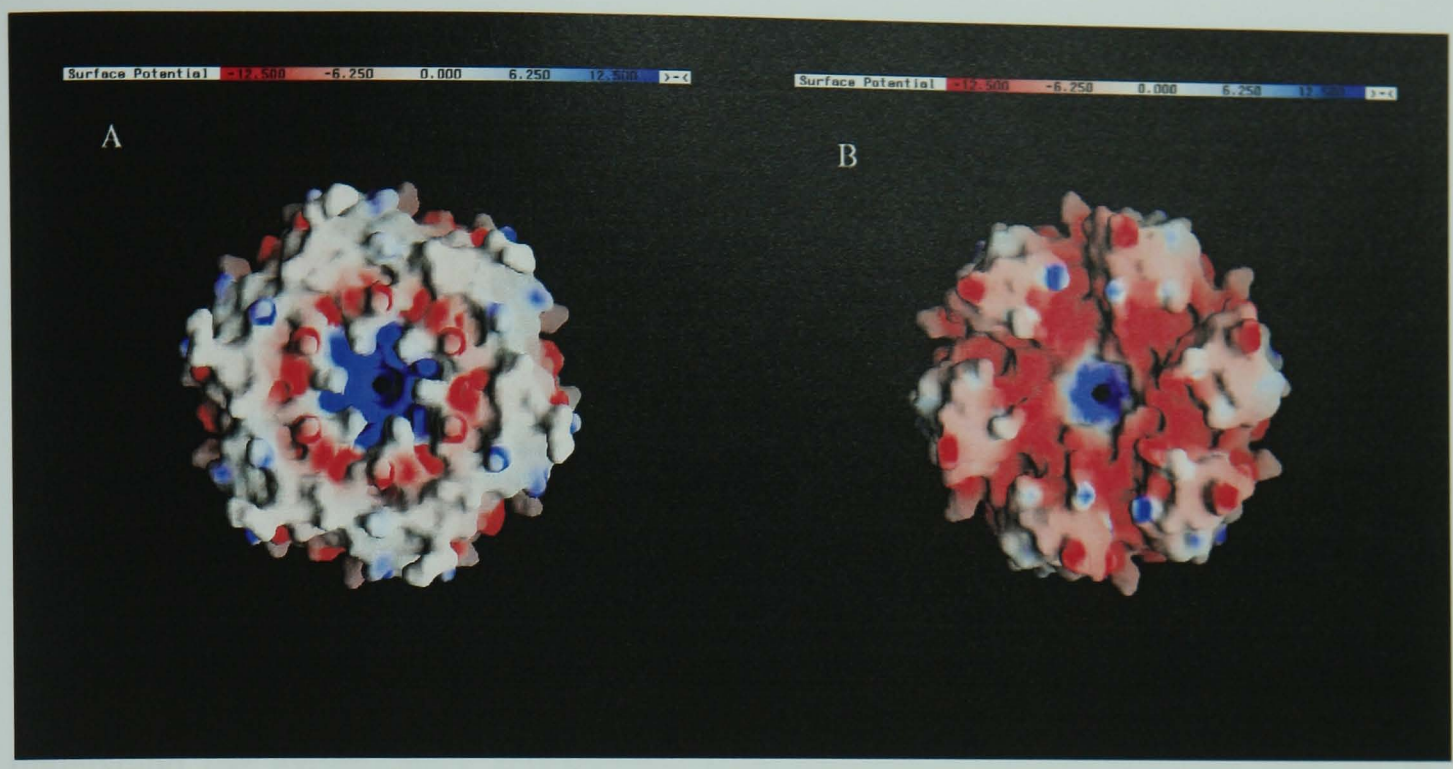
## 6.5 Conclusion

Until now the exact function of Sm–like proteins from archaeons, including AF–Sm2, is unknown. However, based on the structure of the AF–Sm2 hexamer as well as on functional and structural analogy to the well characterised human spliceosomal Sm proteins there can be no doubt about their evolutionary relationship.

In humans seven spliceosomal Sm proteins form the Sm core domain assembling around the Sm site, a U–rich, single stranded stretch of snRNA (Branlant *et al.*, 1982). In the absence of snRNA they form smaller sub–complexes (EFG, $D_1D_2$, $D_3B$) (Plessel *et al.*, 1997) which associate in the presence of the Sm site (Fischer *et al.*, 1985; Feeney *et al.*, 1989; Raker *et al.*, 1996; Raker *et al.*, 1999). Based on the crystal structure of two

complexes, $D_1D_2$ and $D_3B$, as well as on EM studies and biochemical data on the pairwise interaction between individual Sm proteins Kambach *et al.* proposed a model for the human Sm core domain (Raker *et al.*, 1996; Plessel *et al.*, 1997; Camasses *et al.*, 1997; Fury *et al.*, 1997; Kambach *et al.*, 1999). According to this model the seven Sm proteins are arranged in a seven–membered ring interacting with each other the same way as observed in the sub–complex structures. The ring has a positively charged central hole. This central hole is supposed to bind the Sm site of the snRNA, which is wide enough to easily accommodate single stranded RNA. This assumption is supported by experimental evidence provided by UV cross–linking experiments (Heinrichs *et al.*, 1992).

Although there are no similar experimental data available yet for AF–Sm2, the formation of a hexameric complex observed both in solution and in the crystal suggests an analogous role, namely RNA binding for AF–Sm2 as well. The interactions between monomers in the AF–Sm2 hexamer are similar as in the proposed model of the human Sm core domain. The consequence of the fewer number of monomers in the AF–Sm2 ring is a narrower central hole compared to the human Sm core model. Its diameter is ~14 Å if only main chain atoms are considered (compared to ~20 Å in the human core model), but is still wide enough to encompass single stranded RNA. Similarly to the model of the human Sm core domain the central hole is lined with positively charged and polar residues which are capable of interacting well with single stranded RNA (Figure 6.10A & B). In addition, the entrance of the hole on one side of the hexamer is lined with solvent exposed tyrosines (Figure 6.10A,C), whose aromatic rings can form stacking interactions with the RNA bases (for an example see Handa *et al.*, 1999).



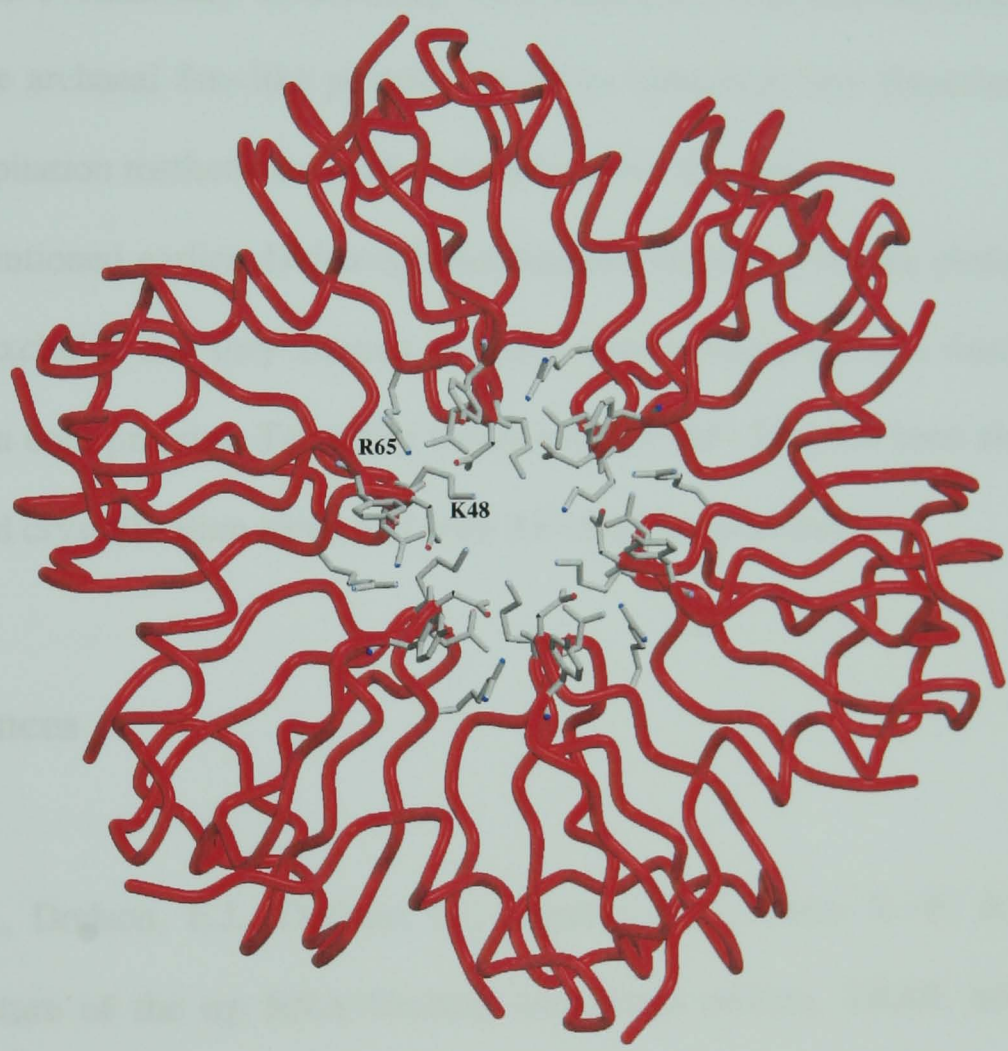




**Figure 6.10 A)** The charge distribution on the solvent accessible surface of one side of the hexamer shows accumulated positive charge in the central hole. The protrusions at the entrance of the hole are tyrosine (Y39) side chains. The figure was made using GRASP (Nicholls *et al.*, 1991). **B)** The same representation showing the other side of the hexamer. **C)** The $C_\alpha$–trace representation of the hexamer in the same orientation showing the

positively charged residues K23 and R65 (labelled) as well as two asparagine side chains (N67 & N68) which could potentially interact with bound single stranded RNA. The figure was prepared with MOLSCRIPT version 2.1.2 (Kraulis, 1991) and RASTER3D version 2.4j (Merritt & Bacon, 1997).

The work presented in the second part of this thesis is mainly of structural nature and might be considered unusual as it deals with the structure of a protein with unknown function. In order to elucidate the function of Sm–like proteins in archaebacteria and to understand the evolutionary relationship with eukaryotic Sm and Sm–like proteins the target of these archaeal Sm–like proteins has to be identified first. Experiments utilising immunoprecipitation methods are under way to achieve this goal.

As mentioned earlier *Archaeoglobus fulgidus* has two Sm–like proteins, therefore it cannot be excluded that they interact and form hetero–oligomers in a similar manner to the human Sm core proteins. To clarify this possibility AF–Sm1 has been also cloned and expressed, and crystallisation alone and with AF–Sm2 is underway.

## 6.6 References

Antson, A.A., Dodson, E.J., Dodson, G., Greaves, R.B., Chen, X.–P. & Gollnick, P. (1999). Structure of the trp RNA–binding attenuation protein, TRAP, bound to RNA. *Nature* **401**, 235–242.

Barton, G.J. (1993). ALSCRIPT a tool to format multiple sequence alignements. *Protein Eng.* **6**, 37–40.

Branlant. C., Krol, A., Ebel, J.–P., Lazar, E., Haendler, B. & Jacob, M. (1982). U2 RNA shares a structural domain with U1, U4 and U5 RNAs. *EMBO J.* **1**, 1259–1265.

Camasses, A., Bragado, N.E., Martin, R., Séraphin, B. & Bordonné, R. (1998). Interactions within the yeast Sm core complex: from proteins to amino acids. *Mol. Cell. Biol.* **18**, 1956–1966.

Cooper, M., Johnston, L.H. & Beggs, J.D. (1995). Identification and characterisation of Uss1p (Sdb23p): a novel U6 snRNA–associated protein with significant similarity to core proteins of small nuclear ribonucleoproteins. *EMBO J.* **14**, 2066–2075.

Feeney, R.J., Suaterer, R.A., Feeney, J.L. & Zieve, G.W. (1989). Cytoplasmic assembly and nuclear accumulation of mature small nuclear ribonucleoprotein particles. *J. Biol. Chem.* **264**, 5776–5783.

Fischer, D.E., Conner, G.E., Reeves, W.H., Wisniewolski, R. & Blobel, G. (1985). Small nuclear ribonucleoprotein particle assembly in vivo: demonstration of a 6S RNA–free core precursor and posttranslational modification. *Cell* **42**, 751–758.

Fury, M.G., Zhang, W., Christodoulopoulos, I. & Zieve, G.W. (1997). Multiple protein:protein interactions between the snRNP common core proteins. *Exp. Cell. Res.* **237**, 63–69.

Handa, N., Nureki, O., Kurimoto, K., Kim, I., Sakamoto, H., Shimura, Y., Muto, Y. & Yokoyama, S. (1999). Structural basis for recognition of the *tra* mRNA precursor by the Sex–lethal protein. *Nature* **398**, 579–585.

Heinrichs, V., Hackl, W. & Lührmann, R. (1992). Direct binding of small nuclear ribonucleoprotein G to the Sm site of small nuclear RNA. Ultraviolet light cross–linking of protein G to the AUU stretch within the Sm site (AAUUUGUGG) of U1 small nuclear ribonucleoprotein reconstituted in vitro. *J. Mol. Biol.* **227**, 15–28.

Hermann, H., Fabrizio, P., Raker, V.A., Foulaki, K., Hornig, H., Brahms, H. & Lührmann, R. (1995). snRNP Sm proteins share two evolutionary conserved sequence motifs which are involved in Sm protein–protein interaction. *EMBO J.* **14**, 2076–2088.

Kambach, C., Walke, S., Young, R., Avis, J.M., de la Fortelle, E., Raker, V.A., Lührmann, R., Li, J. & Nagai, K. (1999). Crystal structure of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs. *Cell* **96**, 375–387.

Kraulis, P.J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950.

Merritt, E.A. & Bacon, D.J. (1997). Raster 3D photorealistic molecular graphics. *Methods in Enzymology* **277**, 505–524.

Nicholls, K.A., Sharp, B. & Honig, B. (1991). Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281–296.

Plessel, G., Lührmann, R. & Kastner, B. (1997). Electron microscopyof assembly intermediates of the snRNP core: morphological similarities between the RNA–free (E.F.G) protein heteromer and the intact snRNP core. *J. Mol. Biol.* **265**, 87–94.

Raker, V.A., Plessel, G. & Lührmann, R. (1996). The snRNP core assembly pathway: identification of stable core protein heteromeric complexes and an snRNP subcore particle *in vitro*. *EMBO J.* **15**, 2256–2269.

Raker, V.A., Hartmuth, K., Kastner, B. & Lührmann, R. (1999). Spliceosomal U snRNP core assembly: Sm proteins assemble onto an Sm site RNA nonanucleotide in a specific and thermodynamically stable manner. *Mol. Cell. Biol.* **19**, 6554–6565.

Salgado–Garrido, J., Bragado–Nilsson, E., Kandels–Lewis, S. & Séraphin, B. (1999). Sm and Sm–like proteins assemble in two related complexes of deep evolutionary origin. *EMBO J.* **18**, 3451–3462.

Séraphin, B. (1995). Sm and Sm–like proteins belong to a large family: identification of proteins of the U6 as well as the U1, U2, U4 and U5 snRNPs. *EMBO J.* **14**, 2089–2098.

# Appendix A

Appendix A shows an alignment of protein sequences found in sequence databases by BLAST and PSI–BLAST (Altschul *et al.*, 1997) based on their sequence homology to S1 nuclease. The sequences of PLC from *Bacillus cereus* and alpha–toxin from *Clostridium perfringens* are not included in the alignment due to their low overall sequence similarity to these proteins. However, the residues responsible for zinc coordination (Figure 3.11) and surprisingly the overall fold (Figure 3.4) are fairly conserved also in PLC and alpha–toxin. The aligned sequences are listed in Table 7.1, the colour coding is explained (according to colprot.par in ClustalX) in the following:

- The rules for assigning colours to the residues. A certain colour is assigned on the basis of a list of consensus (in bold, separated by ':'):

| | |
|---|---|
| glycine = | ORANGE |
| proline = | YELLOW |
| threonine = | GREEN if **t:S:T:%:#** |
| serine = | GREEN if **t:S:T:#** |
| asparagine = | GREEN if **n:N:D** |
| glutamine = | GREEN if **q:Q:E:+:K:R** |
| tryptophane = | BLUE if **%:#:A:C:F:H:I:L:M:V:W:Y:P:p** |
| leucine = | BLUE if **%:#:A:C:F:H:I:L:M:V:W:Y:P:p** |
| valine = | BLUE if **%:#:A:C:F:H:I:L:M:V:W:Y:P:p** |
| isoleucine = | BLUE if **%:#:A:C:F:H:I:L:M:V:W:Y:P:p** |
| methionine = | BLUE if **%:#:A:C:F:H:I:L:M:V:W:Y:P:p** |
| alanine = | BLUE if **%:#:A:C:F:H:I:L:M:V:W:Y:P:p:T:S:s:G** |
| phenylalanine = | BLUE if **%:#:A:C:F:H:I:L:M:V:W:Y:P:p** |
| cysteine = | BLUE if **%:#:A:F:H:I:L:M:V:W:Y:S:P:p** |
| cysteine = | PINK if **C** |
| histidine = | CYAN if **%:#:A:C:F:H:I:L:M:V:W:Y:P:p** |
| tyrosine = | CYAN if **%:#:A:C:F:H:I:L:M:V:W:Y:P:p** |
| glutamic acid = | MAGENTA if **–:D:E:q:Q** |
| aspartic acid = | MAGENTA if **–:D:E:n:N** |
| lysine = | RED if **+:K:R:Q** |
| arginine = | RED if **+:K:R:Q** |

- The explanation of the rules calculating the consensus used for color coding. The underlined characters are single letter amino acid codes; the consensus symbols are shown in bold (as above):

| | | |
|---|---|---|
| **%** = | 60% | w:l:v:i:m:a:f:c:y:h:p |
| **#** = | 80% | w:l:v:i:m:a:f:c:y:h:p |
| **–** = | 50% | e:d |
| **+** = | 60% | k:r |
| **g** = | 50% | g |
| **n** = | 50% | n |
| **q** = | 50% | q:e |
| **p** = | 50% | p |
| **t** = | 50% | t:s |
| **A** = | 85% | a |
| **C** = | 85% | c |
| **D** = | 85% | d |
| **E** = | 85% | e |
| **F** = | 85% | f |
| **G** = | 85% | g |
| **H** = | 85% | h |
| **I** = | 85% | i |
| **K** = | 85% | k |
| **L** = | 85% | l |
| **M** = | 85% | m |
| **N** = | 85% | n |
| **P** = | 85% | p |
| **Q** = | 85% | q |
| **R** = | 85% | r |
| **S** = | 85% | s |
| **T** = | 85% | t |
| **V** = | 85% | v |
| **W** = | 85% | w |
| **Y** = | 85% | y |

| *Alignment ID* | *Protein name* | *Species* | *References* |
|---|---|---|---|
| nuc_s1 | nuclease S1 | *Aspergillus oryzae* | Iwamatsu *et al.*, 1991 |
| nuc_P1 | nuclease P1 | *Penicillinum citrinum* | Maekawa *et al.*, 1991 |
| nuc_Le1 | nuclease Le1 | *Lentinula edodes* | |
| nucI_Hvul | nuclease I | *Hordeum vulgare* | Muramoto *et al.*. 1999 |
| enuc_Hvul | endonuclease | *Hordeum vulgare* | Aoyagi *et al.*. 1998 |
| bfnuc1_Atha | bifunctional nuclease bfn1 | *Arabidopsis thaliana* | |
| sap6 | senescence–associated protein 6 | *Hemerocallis hybrid cultivar* | Panavas *et al.*. 1999 |
| enuc_Zele | endonuclease | *Zinnia elegans* | Aoyagi *et al.*, 1998 |
| CELImenuc_Agra | CEL I mismatch endonuclease | *Apium graveolens* | Yang *et al.*, 2000 |
| bfnuc_Atha | putative bifunctional nuclease | *Arabidopsis thaliana* | |
| bfnuc_Zele1 | bifunctional nuclease | *Zinnia elegans* | Perez–Amador *et al.*, 2000 |
| bfnuc_Zele2 | bifunctional nuclease | *Zinnia elegans* | Perez–Amador *et al.*, 2000 |
| pprot_Atha | putative protein | *Arabidopsis thaliana* | |
| 3'nuc_Ldon | 3'–nucleotidase/ nuclease | *Leishmania donovani* | Debrabant *et al.*, 1995 |
| ssnuc_Lpif | single strand–specific nuclease | *Leishmania pifanoi* | |
| enucs1_Mlot | endonuclease S1 homolog | *Mesorhizobium loti* | Sullivan & Ronson, 1998 |

**Table 7.1** Protein sequences found by BLAST and PSI–BLAST searches based on their sequence similarity to S1 nuclease.


Continued on the next page.

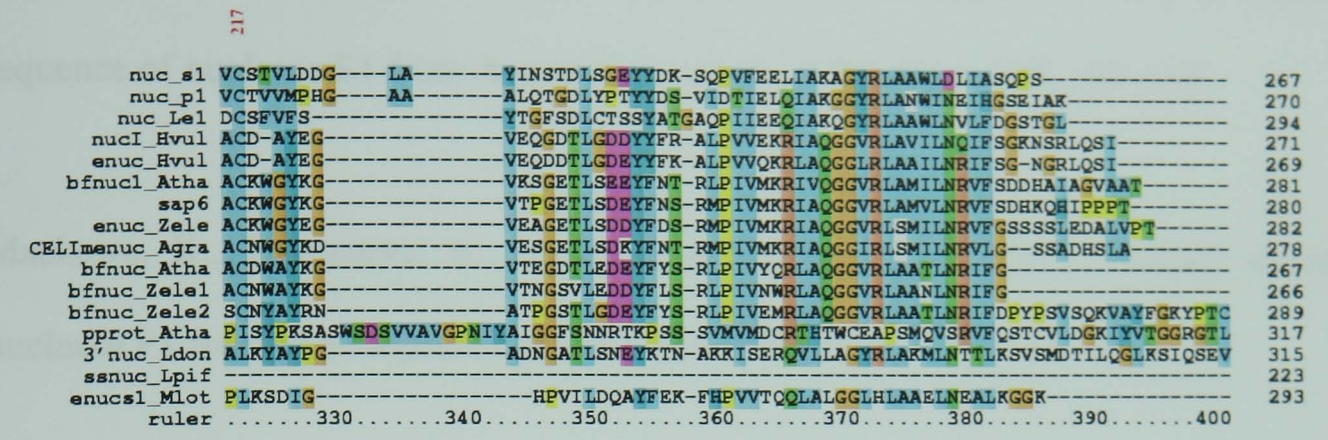



**Figure 7.1** Multiple sequence alignment of protein sequences homologous to S1 nuclease. CLUSTALX version 1.7 was used for the alignment and visualisation (Thompson *et al.*, 1994). Sequence positions thought to be catalytically and structurally important are labelled with red residue numbers corresponding to residue numbers in the S1 nuclease sequence.

## References:

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped BLAST and PSI–BLAST: a new generation of protein database search programs. *Nucleic Acid Res.* **25**, 3389–3402.

Aoyagi, S., Sugiyama, M. & Fukuda, H. (1998). BEN1 and ZEN1 cDNAs encoding S1–type DNases that are associated with programmed cell death in plants. *FEBS Lett.* **429**, 134–138.

Debrabant, A., Gottlieb, M. & Dwyer, D.M. (1995). Isolation and characterisation of the gene encoding the surface membrane 3'–nucleotidase/nuclease of Leishmania donovani. *Mol. Biochem. Parasitol.* **71**, 51–63.

Iwamatsu, A., Aoyama, H., Dibó, G., Tsunasawa, S. & Sakiyama, F. (1991). Amino acid sequence of nuclease S1 from Aspergillus oryzae. *J. Biochem.* **110**, 151–158.

Maekawa, K., Tsunasawa, S., Dibó, G. & Sakiyama, F. (1991). Primary structure of nuclease P1 from Penicillinum citrinum. *Eur. J. Biochem.* **200**, 651–661.

Muramoto, Y., Watanabe., A., Nakamura, T. & Takabe., T. (1999). Enhanced expression of a nuclease gene in leaves of barley plant under salt stress. *Gene* **234**, 315–321.

Panavas, T., Pikula, A., Reid, P.D., Rubinstein, B. & Walker, E.L. (1999). Identification of senescence–associated genes from daylily petals. *Plant Mol. Biol.* **40**, 237–248.

Perez–Amador, M.A. *et al.* (2000). Identification of BFN1, a bifunctional nuclease induced during leaf and stem senescense in arabidopsis. *Plant Physiol.* **122**, 169–179.

Sullivan, J.T. & Ronson, C.W. (1998). Evolution of rhizobia by acquisition of a 500–kb symbiosis island that integrates into a Phe–tRNA gene. *Proc. Natl. Acad. Sci. USA* **95**, 5145–5149.

Thompson, J.D., Higgins, D.G. & Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions–specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.

Yang , B. *et al.* (2000). Purification, cloning and characterisation of the CEL I nuclease. *Biochemistry*, In press.

# Appendix B

**Figure 8.1** The multiple sequence alignment of 91 Sm and Sm–like proteins. The two conserved sequence motifs, *Sm1* and *Sm2* motifs are shown on the top of the alignment. The C–terminal tails after the *Sm2* motif, which vary very much from protein to protein, are not shown in the figure. The colour coding is the same as explained on page 172. Naming conventions:

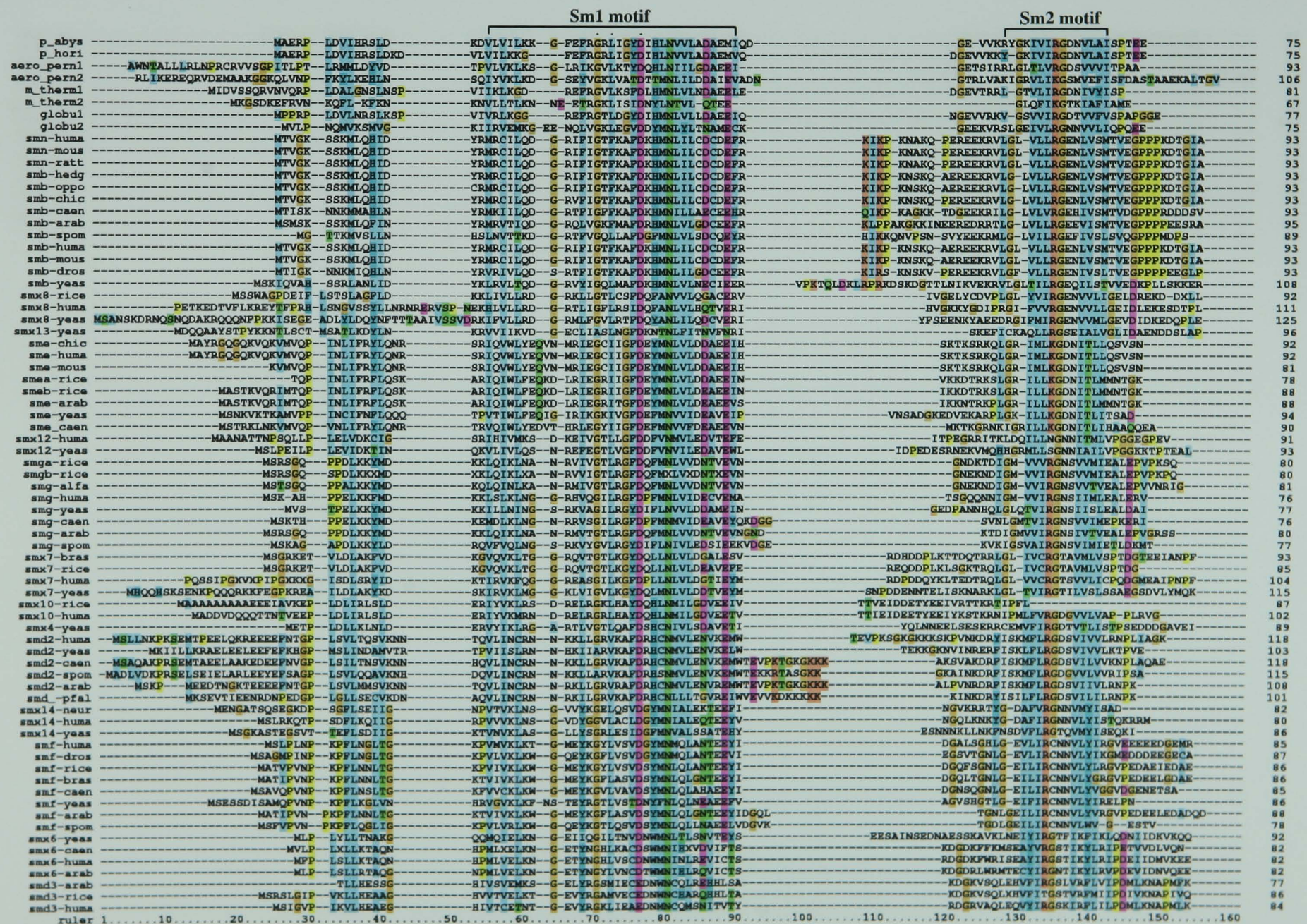
- The first three characters in the label of eukaryotic Sm proteins are either **smb**, **smd**, **sme**, **smf**, **smg** or **smn**.
- For eukaryotic Sm–like proteins the first three characters in the labels are **smx** uniformly. For eukaryotic Sm and Sm–like protein the source organism is indicated by the last four characters of the label as follows:

  **huma,** human; **mous**, mouse; **ratt**, rat; **hedg**, hedgehog; **oppo**, opossum; **chick**, chicken; **caen**, *Ceanorhabditis elegans*; **arab**, *Arabidopsis thaliana*; **spom**, *Schizosaccharomyces pombe*; **dros**, *Drosophila melanogaster*; **yeas**, *Saccharomyces cerevisiae*; **rice**, rice; **alfa**, alfalfa; **bras**, *Brassica campestris pekinensis*; **pfal**, *Plasmodium falsiparum*; **neur**, *Neurospora crassa*; **zeam**, *Zea mais*.

- The archaebacterial Sm–like proteins are placed in the first rows of the alignment. The included sequences are:

  **p_abyss**, *Pyrococcus abyssi*; **p_hori**, *Pyrococcus horikoshii*; **aero_pern1** and **2**, *Aeropyrum pernix*; **m_therm1** and **2**, *Methanobacterium thermoautrotrophicum*; **globu1** and **2**, *Archaeoglobus fulgidus*.

Continued on the next page.



```
smd3-yeas  ----------MTMNGIP--VKLLNEAQG----------HIVSLELTT--G-ATYRGKLVESEDSMNVQLRDVIATE----------PQGAVTHMDQIFVRGSQIKFIVVPDLLKNAPLF--------   84
smd3-spom  ----------MSLC--IKLLHETQG----------HIVTMELEN--G-STYRGKLIEAEDNMNCQMRDISVTA----------RDGR--VSHLDQVYIRGSHIRFLIVPDMLRNAPMFK-------   82
smd3-dros  ----------MSIGVP--IKVLHEAEG----------HIITCETIT--G-EVYRGKLIEAEDNMNCQMTQITVTY----------RDGR--TANLENVYIRGSKIRFLILPDMLKNAPMFK-------   84
smd3-caen  ----------MTSVGVP--IKILHEAEG----------HMVTLETVT--G-EVYRGKLSEAEDNMNCQLAETVVTF----------RDGR--SHQLDNVFIRGNKIRFMILPDMLKNAPMFKNIGRAQK   92
smd1-huma  ----------MKL--VRFLMKLSH----------ETVTIELKN--G-TQVHGTITGVDVSMNTHLKAVKMTL----------KNREPVQLETLSIRGNNIRYFILPDSLPLDTLLV-------   81
smd1-mous  ----------MKL--VRFLMKLSH----------ETVTIELKN--G-TQVHGTITGVDVSMNTHLKAVKMTL----------KNREPVQLETLSIRGNNIRYFILPDSLPLDTLLV-------   81
smd1a-rice ----------MKL--VRFLMKLNN----------ETVTIELKN--G-TVVHGTITGVDISMNTHLKTVKLTL----------KGKNPVTLDHLSVRGNNIRYYILPDSLNL-----------   76
smd1-yeas  ----------MKL--VNFLKKLRN----------EQVTIELKN--G-TTVWGTLQSVSPQMNAILTDVKLTLPQPRLNKLNS--NGIAMASLYLTGGQQPTASDNIASLQYINIRGNTIRQIILPDSLNLDSLLV-------  108
smd1-spom  ----------MKL--VRFLMKLTN----------ETVSIELKN--G-TIVHGTITSVDMQMNTHLKAVKMTV----------KGREPVPVETLSIRGNNIRYYILPDSLPLDTLLI-------   81
smd1-arab  ----------MKL--VRFLMKLNN----------ETVSIELKN--G-TIVHGTITGVDVSMNTHLKAVKLTL----------KGKNPVTLDHLSVRGNNIRYYILPDSLNLETLLV-------   81
smd1-caen  ----------MKL--VRFLMKLSH----------ETVNIELKN--G-TQVSGTIMGVDVAMNTHLRAVSMTV----------KNKEPVKLDTLSIRGNNIRYIILPDPLALDTLLI-------   81
smx5-yeas  ----------MLF--FSFFKTLVD----------QEVVVELKN--D-IEIKGTLQSVDQFLNLKLDNISCTD----------EKKYPHLGSVRNIFIRGSTVRYVYLNKNMVDTNLLQ-------   83
smx5-zeam  ----------MLF--FSYFKELVG----------KEVTVELKN--D-FAIRGTVHSVDQYLNIK----------------------------------------------------------   39
smx5-huma  ----------FFKSLVG----------KDVVVELKN--D-LSICGTLHSVDQYLNIKLTDISVTD----------PEKYPHMLSVKNCFIRGSVVRYVQLPADEVDTQLLQ-------   78
smx9-rice  ----------MASAG---PGLESLVD----------QIISVITND--G-RNIVGTLRGFDQATNIILDESHERV----------YSTREGVQQLVLGLYIIRGDNISVVGEVDEELDARLDL-------   86
smx11-rice ----------MEGGGEEFAIG---------VLISVKTTL--G-EEFEGQIVSFDRPTNLLVIQ----------------------------------------------------------   41
smx1-yeas  ----------MDILKLSDPIG----------NTLIVSLTE--D-RILVGSLVAVDAQMNLLLDHVEERM----------GSSSRMMGLVSVPRRSVKTIMIDKPVLQELTANK-------   80
    ruler 1.....10.....20.....30.....40.....50.....60.....70.....80.....90.....100.....110.....120.....130.....140.....150.....160
```